Universidade de Aveiro Departamento de Matemática
2009

Hilário Amílcar dos Santos Ribeiro Miranda

MÉTODOS ROBUSTOS EM GEOESTATÍSTICA

Universidade de Aveiro Departamento de Matemática
2009

Hilário Amílcar dos Santos Ribeiro Miranda

MÉTODOS ROBUSTOS EM GEOESTATÍSTICA

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Matemática, realizada sob a orientação científica da Doutora Maria Manuela Souto de Miranda, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

Dedico este trabalho aos meus pais pelo seu incansável apoio.

o júri

presidente:

Doutor Eduardo Anselmo Ferreira da Silva
Professor Catedrático da Universidade de Aveiro

vogais:

Doutor João António Branco
Professor Associado do Instituto Superior Técnico da Universidade Técnica de Lisboa

Doutora Andreia Oliveira Hall
Professora Associada da Universidade de Aveiro

Doutora Raquel Menezes Mota Leite
Professora Auxiliar da Escola de Ciências da Universidade do Minho

Doutora Maria Manuela Souto de Miranda
Professora Auxiliar da Universidade de Aveiro (Orientadora)

agradecimentos

Antes de mais, quero deixar um agradecimento sentido à Professora Manuela Souto pelo excelente apoio que me deu e pela sua indispensável orientação durante a realização deste trabalho.
Quero agradecer especialmente à Universidade de Aveiro pelo apoio financeiro e pedagógico que me disponibilizou. Também agradeço à Unidade de Investigação Matemática e Aplicações pelo apoio concedido.
Deixo o meu obrigado para o Professor Ricardo Maronna e para o Professor Manuel Scotto pelas suas valiosas sugestões, que contribuíram para melhorar o teste à estacionaridade da média.
Quero agradecer ao Dr. Jan Keizer por ter cedido o acesso a diversos conjuntos de dados reais e por estar sempre pronto a partilhar o seu conhecimento aplicado da Geoestatística.
Em relação à escrita deste trabalho, tenho que dirigir o meu obrigado ao Professor Paolo Vettori que esteve sempre disponível para ajudar nas dúvidas do LATeX.
Ainda me resta deixar uma palavra a todos aqueles que, de alguma forma, me ajudaram a melhorar este trabalho, como é o caso da Professora Isabel Pereira ou do Professor António Caetano.
Por fim, não posso deixar de agradecer aos meus pais, ao meu irmão e à Josiane por todo o apoio que me deram durante estes últimos anos. Para eles deixo um muito obrigado...

palavras-chave Geoestatística, robustez, variograma, estacionaridade da média, *bootstrap*, estimadores-MM.

resumo

O objectivo principal da presente tese consiste no desenvolvimento de estimadores robustos do variograma com boas propriedades de eficiência.

O variograma é um instrumento fundamental em Geoestatística, pois modela a estrutura de dependência do processo em estudo e influencia decisivamente a predição de novas observações.

Os métodos tradicionais de estimação do variograma não são robustos, ou seja, são sensíveis a pequenos desvios das hipóteses do modelo. Essa questão é importante, pois as propriedades que motivam a aplicação de tais métodos, podem não ser válidas nas vizinhanças do modelo assumido.

O presente trabalho começa por conter uma revisão dos principais conceitos em Geoestatística e da estimação tradicional do variograma. De seguida, resumem-se algumas noções fundamentais sobre robustez estatística. No seguimento, apresenta-se um novo método de estimação do variograma que se designou por estimador de múltiplos variogramas. O método consiste em quatro etapas, nas quais prevalecem, alternadamente, os critérios de robustez ou de eficiência. A partir da amostra inicial, são calculadas, de forma robusta, algumas estimativas pontuais do variograma; com base nessas estimativas pontuais, são estimados os parâmetros do modelo pelo método dos mínimos quadrados; as duas fases anteriores são repetidas, criando um conjunto de múltiplas estimativas da função variograma; por fim, a estimativa final do variograma é definida pela mediana das estimativas obtidas anteriormente. Assim, é possível obter um estimador que tem boas propriedades de robustez e boa eficiência em processos Gaussianos.

A investigação desenvolvida revelou que, quando se usam estimativas discretas na primeira fase da estimação do variograma, existem situações onde a identificabilidade dos parâmetros não está assegurada. Para os modelos de variograma mais comuns, foi possível estabelecer condições, pouco restritivas, que garantem a unicidade de solução na estimação do variograma.

A estimação do variograma supõe sempre a estacionaridade da média do processo. Como é importante que existam procedimentos objectivos para avaliar tal condição, neste trabalho sugere-se um teste para validar essa hipótese.

A estatística do teste é um estimador-MM, cuja distribuição é desconhecida nas condições de dependência assumidas. Tendo em vista a sua aproximação, apresenta-se uma versão do método *bootstrap* adequada ao estudo de observações dependentes de processos espaciais.

Finalmente, o estimador de múltiplos variogramas é avaliado em termos da sua aplicação prática. O trabalho contém um estudo de simulação que confirma as propriedades estabelecidas. Em todos os casos analisados, o estimador de múltiplos variogramas produziu melhores resultados do que as alternativas usuais, tanto para a distribuição assumida, como para distribuições contaminadas.

keywords Geostatistics, robustness, variogram, mean stationarity, MM-estimators, bootstrap.

abstract

The aim of this thesis is to develop robust estimators of the variogram with a high efficiency under Gaussian processes.

The variogram is a fundamental tool in Geostatistics, since it models the process dependence structure and it affects decisively the prediction of unobserved values of the spatial random process.

The usual variogram estimators are not robust in the sense that they are affected by small departures from the model assumptions. That is an important issue, since good properties of the methods may fail in the neighborhoods of the assumed model.

The present work starts with a review of the main concepts on Geostatistics and of the traditional process used in the estimation of the variogram. Also the fundamentals of robust estimation are summarized in the first part of the thesis. Afterwards, a new variogram estimation method is developed, herein called multiple variograms estimator. The method consists of four steps, alternating robustness and efficiency as main criteria. A few robust discrete estimates of the variogram are computed for adjusting the variogram model by least squares. Both procedures are repeated several times, creating multiple variogram estimates. The final estimate is the median of the previous variogram estimates. Thus, the proposal provides an estimator that has good robustness properties and good efficiency for Gaussian processes.

The research revealed that when discrete estimates are used in the first stage of the variogram estimation, the parameters of the model may be unidentifiable. It was possible to establish mild conditions that assure the identifiability of the parameters of the most popular variogram models.

The variogram estimation is always conducted under the assumption that the process mean is stationary. Objective methods that can evaluate that assumption are of extreme importance. Hence, this work recommends a statistical test for verifying the mean stationarity condition. The statistic that is used in the test is an MM-estimator, which has an unknown distribution under the assumed dependence structure. For approximating the distribution of the estimator, an adequate spatial bootstrap version is proposed which can deal with spatial dependent observations.

Finally, the multiple variograms estimator was evaluated from a practical point of view. The work includes a simulation study that confirmed the established properties. In all the considered cases, the multiple variograms estimator performed better than the usual variogram estimators, either considering the assumed distribution or the contaminated distributions.

Índice

Lista de Tabelas

Lista de Figuras

Introdução

A Geoestatística é um ramo da Estatística que se destina ao estudo da distribuição espacial de grandezas que representam propriedades de recursos naturais, como por exemplo, recursos geológicos, hidrológicos e ecológicos. A modelação dos fenómenos é efectuada com base em processos estocásticos espaciais.

O que distingue um processo geoestatístico dos restantes processos espaciais, como por exemplo os processos discretos (*lattice data*), ou os padrões espaciais (*spatial patterns*), é o facto de, nos processos geoestatísticos, se considerar que as localizações onde se obtêm as realizações do processo são fixas (não aleatórias) e ainda que o processo tem índice espacial contínuo. A suposição de que o processo tem índice espacial contínuo significa que, entre duas quaisquer localizações onde existem realizações do processo, é sempre possível (teoricamente) obter um conjunto infinito de outras localizações onde o processo é observável. Consequentemente, as localizações podem variar continuamente ao longo da região que se pretende estudar. Note-se que a noção de continuidade que foi referida está associada à região onde o processo é estudado e não ao atributo em estudo. O atributo pode ser uma variável aleatória discreta ou contínua, conforme a natureza do problema em questão.

A Geoestatística surgiu nos finais da década dos anos cinquenta e veio dar resposta à necessidade de modelação de recursos geológicos, como por exemplo, a caracterização da concentração de metais em jazigos minerais ou o estudo da qualidade de águas subterrâneas. Daí surgiu o prefixo *geo*, que foi pela primeira vez utilizado por Hart (1954). Poucos anos depois, G. Matheron deu um contributo decisivo para a orientação actual, com a publicação de alguns artigos (Matheron (1962), Matheron (1963) e Matheron (1971)). Inicialmente, a Geoestatística era uma disciplina híbrida, que englobava um pouco de Engenharia de Minas, de Geologia, de Matemática e de Estatística. A partir

do contributo de Matheron, a Geoestatística ganhou o reconhecimento dos geólogos, que encontraram nos métodos geoestatísticos uma forma fácil de representar graficamente as propriedades do solo de um terreno inteiro, a partir de uma amostra com poucas observações.

Actualmente, a Geoestatística é utilizada pelas mais diversas ciências da terra, do ambiente e até da saúde. O interesse que os vários ramos da ciência encontram nos métodos geoestatísticos, resulta da sua capacidade única para modelar um processo estocástico que se observa em localizações fixas e que se pretende estudar numa região com índice espacial contínuo.

Apesar da Geoestatística ser uma disciplina bem sistematizada, a investigação de métodos geoestatísticos continua a ser um tópico de grande actualidade. Por um lado, o forte desenvolvimento que se tem verificado nos meios informáticos tem aberto novas oportunidades de investigação. Os recursos informáticos actuais permitem estudar e validar procedimentos que antigamente estavam completamente fora de alcance. Por outro lado, a evolução da ciência em geral, conduziu a novas aplicações dos métodos geoestatísticos a novas áreas do saber e a novos desafios. Como sintoma da evolução em curso e da utilização cada vez mais vasta da Geoestatística, refira-se que uma das principais revistas científicas da área, a *Mathematical Geology*, alterou o título em Janeiro de 2008 para *Mathematical Geosciences*.

Em Geoestatística existem duas áreas de trabalho fundamentais, a obtenção do variograma e a *krigagem*. O variograma é o instrumento que permite caracterizar a estrutura de dependência existente nas observações do processo. É um instrumento muito importante, pois vai influenciar decisivamente a fase de *krigagem*. A designação de *krigagem* foi introduzida por G. Matheron, em honra de D. G. Krige, para designar um conjunto de métodos de predição de futuras observações do processo.

O presente trabalho incide sobre a estimação robusta do variograma. Tradicionalmente, o variograma é estimado em duas fases distintas. A primeira fase consiste em estimar pontualmente o variograma a partir da amostra do processo. Obtém-se assim um conjunto discreto de estimativas do variograma. No entanto, como a região em estudo tem índice espacial contínuo, esse conjunto de estimativas não é suficiente para caracterizar a estrutura de dependência entre quaisquer variáveis aleatórias do

processo. Nesse sentido, é necessário passar a uma segunda etapa da estimação, que consiste em encontrar um modelo de variograma que se aproxime o mais possível das estimativas pontuais obtidas na primeira fase de estimação.

Os métodos tradicionais de estimação do variograma têm boas propriedades, no entanto, não são robustos – isto significa que pequenos desvios das hipóteses assumidas nos modelos podem provocar grandes alterações nos resultados. Por isso, surge a necessidade de investigar e desenvolver métodos geoestatísticos robustos, capazes de suportar pequenas alterações das hipóteses do modelo, sem que produzam maus resultados.

A necessidade de métodos robustos já há muito que era sentida por vários autores quando Box (1953) introduziu, pela primeira vez, o termo *robustez*. No entanto, foram os trabalhos de Huber (1964) e de Hampel (1968) que deram um impulso decisivo à teoria da robustez.

Hoje em dia, é consensual a necessidade de recorrer a técnicas robustas e há um esforço considerável para encontrar procedimentos que constituam alternativas robustas para a análise estatística convencional, de acordo com os diversos modelos.

A importância da robustez estatística na estimação do variograma já foi evidenciada em Cressie e Hawkins (1980). Cressie e Hawkins notaram que os estimadores tradicionais do variograma podem ser bastante afectados por pequenas alterações das hipótese do modelo. Numa tentativa de diminuir o problema, os autores propuseram uma modificação do estimador usual que, no entanto, também não goza de robustez estatística, tal como ela é entendida actualmente.

Apesar da preocupação já manifestada no artigo anteriormente referido, a robustez em Geoestatística tem sido relativamente pouco investigada. Uma excepção é o trabalho desenvolvido em Genton (1998a). Nesse artigo, o autor propôs que o variograma fosse pontualmente estimado através de um estimador robusto, apresentado por Rousseeuw e Croux (1993) noutro contexto. De facto, este último estimador, generalizadamente conhecido por Q_n, consegue resistir bem a afastamentos das hipóteses do modelo, mas conduz a uma grande perda de eficiência durante a segunda etapa de estimação do variograma.

Por outro lado, os principais métodos robustos foram desenvolvidos para cenários de observações independentes e identicamente distribuídas (*i.i.d.*). Por isso, o seu desempenho e as suas propriedades estão pouco exploradas em modelos onde se pretende assumir a dependência entre as observações, como é o caso da Geoestatística. Surge assim a necessidade de investigar procedimentos de estimação do variograma que sejam robustos, sem perder muita eficiência em relação aos estimadores tradicionais.

Assim, o objectivo principal deste trabalho consiste no desenvolvimento de um método de estimação do variograma que consiga conciliar boas propriedades de robustez com boa eficiência em modelos normais.

Do estudo efectuado, resulta a proposta de um método de estimação do variograma, também composto por várias etapas, que desde já se passa a descrever em linhas gerais. O estimador resultante será designado por *estimador de múltiplos variogramas*.

Na primeira fase, estima-se pontualmente o variograma a partir da amostra do processo, usando um método robusto, altamente resistente; o número de estimativas é reduzido, tirando proveito de um resultado de Lahiri, Lee e Cressie (2002) que relaciona o número de estimativas pontuais do variograma com a eficiência do método dos mínimos quadrados. De seguida, utiliza-se o conjunto de estimativas pontuais anteriormente obtidas para estimar os parâmetros do modelo de variograma; nesta etapa, o critério que prevalece é o da eficiência, com recurso ao método dos mínimos quadrados. As duas fases anteriores conduzem à obtenção de uma única estimativa do variograma, o que dá um panorama muito incompleto do modelo. Para além disso, é especialmente importante a forma do variograma junto à origem, o que requer sensibilidade das técnicas a utilizar. Por isso, considerou-se que a aplicação de métodos robustos devia ser sustentada por uma quantidade de observações que facilitasse a identificação dos pontos concordantes com o modelo. Assim, tendo em conta a continuidade do domínio, repetem-se os procedimentos anteriores, fazendo variar os pontos onde se obtêm as estimativas pontuais do variograma. Deste modo, é possível calcular múltiplas estimativas dos parâmetros do modelo de variograma, a partir da mesma amostra do processo. Esta etapa da estimação é muito importante porque permite aumentar a eficiência do método. Finalmente, partindo do conjunto de estimativas dos parâmetros encontradas na fase anterior, determinam-se as estimativas centrais de cada parâmetro do modelo

de variograma. Neste caso, as estimativas centrais são determinadas através da mediana. Deste modo, a última fase volta a reflectir preocupações de robustez, uma vez que a eficiência já foi assegurada anteriormente. São estas últimas estimativas que definem a estimativa final da curva do variograma.

O trabalho foi desenvolvido assumindo algumas condições indispensáveis de estacionaridade. No mínimo, é necessário supor que os processos geoestatísticos são intrinsecamente estacionários, embora alguns resultados suponham condições de estacionaridade de segunda ordem. Quando a média do processo existe, é fundamental assumir que ela é constante em qualquer uma das condições de estacionaridade referidas. Ora, é usual que esta hipótese seja assumida sem que existam procedimentos formais para a testar. A aceitação costuma ser baseada apenas numa análise preliminar de dados. Quando o investigador suspeita que a média não é constante, recorre a transformações de dados para repor a estacionaridade da média e prosseguir com a análise estatística. Porém, este procedimento é sempre subjectivo, pois depende da opinião do investigador. Assim, neste trabalho propõe-se um teste estatístico que permite decidir se um processo geoestatístico apresenta estacionaridade da média.

Para efectuar o referido teste sugere-se um método de projecções ortogonais, que reequaciona o teste à estacionaridade da média na forma de um problema de regressão. Como estatística do teste, optou-se por utilizar um estimador-MM, uma vez que este tipo de estimadores consegue conciliar boas propriedades de robustez e de eficiência em modelos normais. No entanto, a distribuição destes estimadores não é conhecida sob condições de dependência. Sendo assim, foi necessário aproximar a sua distribuição e, para isso, recorreu-se à metodologia *bootstrap*, especificamente adaptada.

Uma vez que existe dependência probabilística entre as observações do processo geoestatístico, a adaptação que se considerou foi inspirada no *bootstrap* por blocos. O *bootstrap* por blocos tem sido utilizado por diversos autores, como por exemplo Hall (1985) ou Lahiri (2003), uma vez que a introdução de blocos no processo de reamostragem permite conservar a estrutura de dependência dentro de cada um dos blocos. O que distingue o *bootstrap* que aqui se utiliza para aproximar a distribuição da estatística do teste, dos procedimentos já divulgados na literatura, é o facto de se associar o *bootstrap* ao método das projecções. Esta técnica permite aproximar a

distribuição do estimador-MM e, assim, decidir objectivamente se existem motivos para rejeitar a condição de estacionaridade da média.

A avaliação do estimador de múltiplos variogramas foi efectuada de várias formas. Em primeiro lugar verificou-se, formalmente, que este estimador tem boas propriedades. É um estimador consistente, robusto e tem uma distribuição assintótica normal. Seguidamente, comparou-se o comportamento do estimador de múltiplos variogramas com o comportamento de outros estimadores que são frequentemente utilizados nas aplicações. Para efectuar a comparação, começou por se fazer um estudo de simulação detalhado, o qual revelou bons resultados. Em todos os processos contemplados, quer com amostras sem contaminação, quer com amostras contaminadas, o estimador proposto revelou um desempenho melhor do que o dos restantes estimadores considerados. Por fim, ainda se estudou o comportamento do estimador com um conjunto de dados reais, já publicados na literatura. Neste caso, os resultados obtidos pelo estimador de múltiplos variogramas encontram-se próximos dos resultados devolvidos pelo outro estimador robusto utilizado.

No decorrer do estudo das propriedades do estimador de múltiplos variogramas verificou-se que, ao usar estimativas pontuais do variograma na primeira fase do processo, os parâmetros do modelo podem não ser identificáveis. Esse aspecto tem sido ignorado na literatura e é de grande relevância prática, pois pode resultar no cálculo de estimativas indesejáveis, erradamente atribuídas a dificuldades numéricas e computacionais. Ao estudar o problema, foi possível estabelecer condições de identificabilidade para os modelos de variograma mais comuns nas aplicações, condições essas que também são válidas quando se usa o estimador tradicional.

É de salientar o papel fundamental dos meios informáticos, sem os quais seria inviável todo o trabalho de cálculo que foi efectuado. Os métodos tradicionais de estimação do variograma envolvem um volume considerável de cálculos e encontram-se disponíveis em diversos programas estatísticos. Um desses programas é o programa *R*, que disponibiliza a *package geoR*, a qual foi desenvolvida por Ribeiro Jr. e Diggle (2001) (para mais informações sobre o programa *R* deve-se consultar R Development Core Team (2008)). Ao longo deste trabalho utiliza-se o programa *R* de base e ainda

algumas das suas *packages*. Para efectuar o cálculo das estimativas, utilizam-se comandos simples que, por sua vez, recorrem a funções já programadas e devidamente testadas. Todas as funções, bem como o *software R*, são de acesso livre e, por isso, qualquer investigador interessado os pode utilizar.

Uma vez apresentadas as ideias gerais que irão ser desenvolvidas ao longo do trabalho, resume-se a estrutura desta dissertação.

No **Capítulo 1** faz-se uma caracterização geral dos processos geoestatísticos unidimensionais e apresentam-se os conceitos essenciais em Geoestatística. De entre esses conceitos, salientam-se as condições de estacionaridade e a definição de variograma.

A metodologia tradicional de estimação do variograma é apresentada ao longo do **Capítulo 2**. Para além de se descreverem os métodos usuais de estimação, também se referem as suas propriedades mais importantes.

O **Capítulo 3** é dedicado à robustez. Definem-se os conceitos básicos da robustez estatística e apresentam-se alguns estimadores de localização, de escala e de regressão, os quais são actualmente recomendados devido às suas propriedades.

Depois de enquadrado o tema e de revistas as noções fundamentais, o trabalho original é descrito a partir do **Capítulo 4**. No **Capítulo 4** faz-se uma revisão das várias versões *bootstrap* existentes, quer para observações *i.i.d.*, quer para observações de estruturas temporais. Seguidamente, propõe-se um método de reamostragem espacial por blocos circulares.

Ao longo do **Capítulo 5** estuda-se a estacionaridade da média de um processo geoestatístico, apresentando um procedimento para testar essa hipótese. Também se investigam as distribuições assintóticas do estimador de mínimos desvios absolutos e dos estimadores-MM sob condições de dependência. O capítulo termina com uma aplicação do teste a um conjunto de dados reais.

No **Capítulo 6** estuda-se, detalhadamente, a estimação robusta do variograma. Em primeiro lugar, revêem-se as principais propostas robustas de estimação do variograma; seguidamente, apresenta-se o novo método de estimação. Este capítulo inclui ainda a demonstração de resultados, os quais estabelecem condições para garantir que

a estimação do variograma tenha unicidade de solução. São ainda estudadas as propriedades assintóticas do estimador de múltiplos variogramas.

No **Capítulo 7** avalia-se o desempenho do estimador de múltiplos variogramas em aplicações, comparando-o com o dos estimadores usuais. Apresenta-se um estudo de simulação e uma aplicação a um conjunto de dados reais.

Para finalizar, nas **Conclusões** resumem-se os resultados encontrados e identificam-se tópicos com interesse para trabalho futuro.

Capítulo 1

Processos geoestatísticos unidimensionais

No presente capítulo apresentam-se vários conceitos essenciais em Geoestatística. A secção **1.1** contém uma caracterização geral dos processos geoestatísticos unidimensionais. Tal como acontece com outros tipos de processos estocásticos, para que se possa efectuar o estudo dos processos geoestatísticos, é necessário assumir que existem algumas condições de estabilidade na distribuição desses processos. Tais condições são expressas através do conceito de estacionaridade, em relação ao qual se resumem as noções mais relevantes na secção **1.2**. Finalmente, a secção **1.3** contém matéria mais específica e fundamental para o desenvolvimento do trabalho, incluindo a definição de variograma e as suas propriedades.

1.1 Caracterização do processo

Um processo geoestatístico $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$ é um processo aleatório em que $\mathbf{s}$ representa um ponto que se desloca continuamente sobre uma região D. Esta região D, que é designada por domínio do processo geoestatístico, é definida como sendo um subconjunto de $\mathbb{R}^d$ (com $d \in \mathbb{N}$), fixo e com volume positivo. Em terminologia geoestatística, dá-se o nome de localização ao ponto $\mathbf{s}$, uma vez que, nas aplicações, $\mathbf{s}$ representa o local onde se observa a realização da variável aleatória $Z(\mathbf{s})$ que é objecto de estudo. A principal característica que distingue os processos geoestatísticos dos outros processos espaciais é o facto da localização $\mathbf{s}$ variar continuamente ao longo do domínio $D \subseteq \mathbb{R}^d$. Note-se que a noção de continuidade referida se encontra associada ao domínio e não

à variável aleatória $Z(\mathbf{s})$. A variável aleatória pode não ser de natureza contínua.

Um processo $Z(\mathbf{s})$ é geralmente caracterizado em dimensão finita, através da sua função de distribuição de probabilidade (*f.d.p.*) conjunta, definida para um número finito de localizações,

$$F_{\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_n}(z_1, z_2, ..., z_n) = \mathrm{P}[Z(\mathbf{s}_1) \leq z_1 \wedge Z(\mathbf{s}_2) \leq z_2 \wedge ... \wedge Z(\mathbf{s}_n) \leq z_n],$$

onde $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n$ são localizações do domínio D, $z_1, z_2, ..., z_n$ pertencem a $\mathbb{R}$ e n é um número natural qualquer. Esta função satisfaz as condições de Kolmogorov:

- F mantém-se invariante sempre que o par $(\mathbf{s}_i, z_i), i = 1, ..., n$, for sujeito à mesma permutação;

- $F_{\mathbf{s}_1,...,\mathbf{s}_n,\mathbf{s}_{n+1}...,\mathbf{s}_{n+m}}(z_1, ..., z_n, +\infty, ..., +\infty) = F_{\mathbf{s}_1,...,\mathbf{s}_n}(z_1, ..., z_n)$ para qualquer número natural m.

Os processos geoestatísticos mais usuais têm domínios contidos em $\mathbb{R}^2$ ou em $\mathbb{R}^3$, uma vez que as variáveis aleatórias mais estudadas se observam no plano ou no espaço. Por exemplo, quando se pretende estudar a densidade populacional de uma dada região, esta poderá ser representada através de um processo geoestatístico $Z(\mathbf{s})$ cujo domínio D é um conjunto de pares ordenados $(\text{latitude}, \text{longitude}) \in \mathbb{R}^2$. Como exemplo de um processo com domínio em $\mathbb{R}^3$, suponha-se que se pretende averiguar qual é a concentração $Z(\mathbf{s})$ de um dado metal num determinado subsolo; o processo geoestatístico poderá ter, como domínio D, um conjunto de ternos ordenados $(\text{latitude}, \text{longitude}, \text{profundidade})$, *i.e.*, $D \subset \mathbb{R}^3$. É de salientar que, como as coordenadas geográficas dos exemplos anteriores se encontram expressas em termos de latitude e longitude, a distância euclidiana não é adequada para medir as distâncias entre as localizações, principalmente quando o domínio que se pretende estudar é grande. Nesses casos, recomenda-se a utilização da distância geodésica, a qual tem em conta a curvatura do globo terrestre para medir as distâncias.

Como os exemplos do parágrafo anterior deixam antever, para quaisquer duas localizações distintas $\mathbf{s}_i$ e $\mathbf{s}_j$ do domínio D, as variáveis $Z(\mathbf{s}_i)$ e $Z(\mathbf{s}_j)$ não são, em geral, independentes e podem mesmo não ser identicamente distribuídas.

1.2 Condições de estacionaridade

As condições de estacionaridade são hipóteses que se colocam sobre os processos aleatórios para garantir que estes gozam de certas propriedades fundamentais em todo o seu domínio. Estas condições garantem uma regularidade nas variáveis aleatórias, a qual é essencial para que se possam utilizar posteriormente métodos da inferência estatística.

Existem diferentes condições de estacionaridade que se passam a apresentar.

Definição 1.2.1. Considere-se um processo geoestatístico $Z(\mathbf{s})$ com domínio $D \subseteq \mathbb{R}^d$ e com *f.d.p.* conjunta $F_{\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_n}(z_1, z_2, ..., z_n)$, onde $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n \in D$, $z_1, z_2, ..., z_n \in \mathbb{R}$ e $n \in \mathbb{N}$.

Diz-se que $Z(\mathbf{s})$ apresenta *estacionaridade forte*, ou *estrita*, se para qualquer conjunto finito de localizações $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$, para qualquer conjunto finito de números reais $\{z_1, z_2, ..., z_n\}$ e para qualquer vector $\mathbf{h} \in \mathbb{R}^d$ tal que $\mathbf{s}_i + \mathbf{h} \in D$ $(i = 1, ..., n)$,

$$F_{\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_n}(z_1, z_2, ..., z_n) = F_{\mathbf{s}_1+\mathbf{h},\mathbf{s}_2+\mathbf{h},...,\mathbf{s}_n+\mathbf{h}}(z_1, z_2, ..., z_n), \quad \forall_{n\in\mathbb{N}}.$$

$\Diamond$

Quando um processo verifica a estacionaridade forte, ele mantém-se invariante sempre que se aplica uma translação a um qualquer conjunto das suas localizações. Deste modo, a *f.d.p.* $F_{\mathbf{s}_1,\mathbf{s}_2,...,\mathbf{s}_n}$ apenas depende da posição relativa entre as localizações e é independente do ponto do domínio onde essas localizações se encontram. Isto implica que as variáveis aleatórias do processo sejam identicamente distribuídas.

Tal como o próprio nome indica, esta condição de estacionaridade é bastante forte para ser assumida em diversos fenómenos naturais. Por isso, utilizam-se outros conceitos de estacionaridade, menos restritivos.

Uma das alternativas é a estacionaridade de segunda ordem, também chamada estacionaridade fraca ou em sentido lato. Quando se utiliza este conceito assume-se que, para qualquer $\mathbf{s} \in D$, $Z(\mathbf{s})$ tem momentos de segunda ordem finitos.

Definição 1.2.2. Um processo geoestatístico $Z(\mathbf{s})$ com domínio $D \subseteq \mathbb{R}^d$ diz-se *estacionário de segunda ordem*, com *estacionaridade fraca* ou *estacionário em sentido lato*, quando verifica as seguintes condições:

$$\forall_{\mathbf{s}\in D} \quad \mathrm{E}[Z(\mathbf{s})] = \mu(\mathbf{s}) = \mu \in \mathbb{R}; \tag{1.2.1}$$

$$\forall_{\mathbf{s}_i,\mathbf{s}_j \in D} \quad \mathrm{Cov}[Z(\mathbf{s}_i), Z(\mathbf{s}_j)] = C(\mathbf{s}_i - \mathbf{s}_j). \tag{1.2.2}$$

$\Diamond$

A condição (1.2.1) garante que as variáveis aleatórias do processo têm todas a mesma esperança μ. A propriedade do valor esperado do processo ser constante será designada por *estacionaridade da média*. Quando não se verifica a estacionaridade da média, a função $\mu(\mathbf{s})$ chama-se a *tendência* do processo $Z(\mathbf{s})$.

A segunda condição da **Definição 1.2.2** garante que a covariância entre duas quaisquer variáveis aleatórias do processo, depende apenas da diferença entre as suas localizações. A função C definida em (1.2.2) chama-se o *covariograma* do processo, ou a *função de covariância estacionária*.

Apesar da estacionaridade de segunda ordem também ser designada por estacionaridade fraca, existe ainda um outro tipo de estacionaridade que impõe menos restrições ao processo. É a denominada estacionaridade intrínseca.

Definição 1.2.3. Um processo geoestatístico $Z(\mathbf{s})$ de domínio $D \subseteq \mathbb{R}^d$ apresenta *estacionaridade intrínseca* se

$$\forall_{\mathbf{s}_i,\mathbf{s}_j \in D} \quad \mathrm{E}[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] = 0; \tag{1.2.3}$$

$$\forall_{\mathbf{s}_i,\mathbf{s}_j \in D} \quad \mathrm{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] = 2\gamma(\mathbf{s}_i - \mathbf{s}_j). \tag{1.2.4}$$

$\Diamond$

As variáveis aleatórias $Z(\mathbf{s}_i) - Z(\mathbf{s}_j), \forall_{\mathbf{s}_i,\mathbf{s}_j \in D}$, chamam-se os *incrementos* do processo.

Caso $Z(\mathbf{s})$ tenha esperança finita, a condição (1.2.3) é apenas uma forma diferente de representar a estacionaridade da média afirmada em (1.2.1). A diferença reside no facto da estacionaridade da média ser agora expressa em termos dos incrementos. Ao longo deste trabalho, considera-se que o processo tem momentos de primeira e segunda ordem finitos.

A segunda condição da **Definição 1.2.3** garante que a variância da diferença entre duas quaisquer variáveis aleatórias do processo, é função do vector definido pela diferença entre as suas localizações; por outras palavras, a variância de um incremento apenas depende do vector que separa as variáveis aleatórias que o definem.

A função 2γ definida em (1.2.4) chama-se o *variograma* do processo e será focada com detalhe na secção seguinte. A função γ designa-se por *semivariograma* do processo.

No que diz respeito às relações existentes entre os vários tipos de estacionaridade, é fácil verificar que, quando $Z(\mathbf{s})$ tem variância finita, então a estacionaridade forte implica a estacionaridade de segunda ordem.

Por outro lado, tendo em atenção que, para qualquer $\mathbf{s}_i, \mathbf{s}_j \in D$,

$$\mathrm{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] = \mathrm{Var}[Z(\mathbf{s}_i)] + \mathrm{Var}[Z(\mathbf{s}_j)] - 2\mathrm{Cov}[Z(\mathbf{s}_i), Z(\mathbf{s}_j)],$$

se existir o covariograma, verifica-se que

$$2\gamma(\mathbf{s}_i - \mathbf{s}_j) = 2C(\mathbf{0}) - 2C(\mathbf{s}_i - \mathbf{s}_j), \tag{1.2.5}$$

uma vez que $\mathrm{Var}[Z(\mathbf{s}_i)] = \mathrm{Var}[Z(\mathbf{s}_j)] = C(\mathbf{0})$.

A equação (1.2.5) garante que, se existir estacionaridade de segunda ordem, então também existe o variograma, e assim também existe estacionaridade intrínseca. Consequentemente, a estacionaridade forte implica a estacionaridade de segunda ordem que, por sua vez, implica a estacionaridade intrínseca.

As implicações recíprocas nem sempre se verificam. No entanto, se o processo geoestatístico for Gaussiano, ou seja, se para qualquer conjunto de localizações $\mathbf{s}_1, ..., \mathbf{s}_n \in D$, $(Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))$ seguir uma distribuição conjunta normal, então a estacionaridade de segunda ordem passa a ser equivalente à estacionaridade forte. Isto acontece porque a distribuição normal fica completamente definida pelo vector de médias e pela matriz de covariâncias. Quanto à equivalência entre a estacionaridade intrínseca e a estacionaridade de segunda ordem, note-se que, quando existe estacionaridade intrínseca (existe o variograma) e quando existe e é constante a variância do processo $Z(\mathbf{s})$, $\mathrm{Var}[Z(\mathbf{s})] = C(\mathbf{0})$, então, por (1.2.5), também existe a quantidade $2C(\mathbf{0}) - 2\gamma(\mathbf{s}_i - \mathbf{s}_j) = 2C(\mathbf{s}_i - \mathbf{s}_j)$ que define o covariograma. Logo, a existência de variância constante $C(\mathbf{0})$ assegura a equivalência entre os dois tipos de estacionaridade.

1.3 A importância do variograma

O variograma, tal como o covariograma, é uma função fundamental em Geoestatística, uma vez que modela a estrutura de dependência do processo. No entanto, em Geoestatística, o variograma assume um papel muito mais relevante do que o covariograma. Ao longo da presente secção, apresentam-se as definições de ambas as funções, comparando as suas propriedades e realçando as vantagens do variograma em relação ao covariograma.

1.3.1 Definições e principais propriedades

Seja $Z(\mathbf{s})$ um processo geoestatístico que se admite ser estacionário de segunda ordem. Como se referiu anteriormente, esta condição de estacionaridade garante que a covariância entre duas realizações do processo depende apenas da diferença entre as localizações. Começa-se por apresentar a definição da função covariograma.

Definição 1.3.1. Para qualquer localização $\mathbf{s} \in D$ e para qualquer vector $\mathbf{h} \in \mathbb{R}^d$ tal que $\mathbf{s} + \mathbf{h} \in D$, chama-se *covariograma* à função

$$\begin{array}{rccl} C\colon & \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ & \mathbf{h} & \longmapsto & C(\mathbf{h}) = \mathrm{Cov}[Z(\mathbf{s}), Z(\mathbf{s}+\mathbf{h})]. \end{array}$$

$\Diamond$

O covariograma diz-se *isotrópico* quando depende do vector $\mathbf{h}$ apenas através da sua norma. Assim, um covariograma isotrópico não depende da direcção de $\mathbf{h}$. Então, para qualquer localização $\mathbf{s}$, a covariância entre $Z(\mathbf{s})$ e qualquer variável localizada sobre uma circunferência de raio $\|\mathbf{h}\|$ é constante.

Em fenómenos naturais, quanto mais próximas estão duas localizações $\mathbf{s}_i$ e $\mathbf{s}_j$, mais se espera que as realizações $Z(\mathbf{s}_i)$ e $Z(\mathbf{s}_j)$ sejam correlacionadas. Por isso, em grande parte dos casos espera-se que, para uma direcção fixa, o valor do covariograma $C(\mathbf{s}_i - \mathbf{s}_j)$ diminua quando $\|\mathbf{s}_i - \mathbf{s}_j\|$ aumenta. A Figura 1.1 mostra três exemplos de covariogramas isotrópicos.

Para que uma função C seja um covariograma, tem que satisfazer algumas propriedades, tais como:

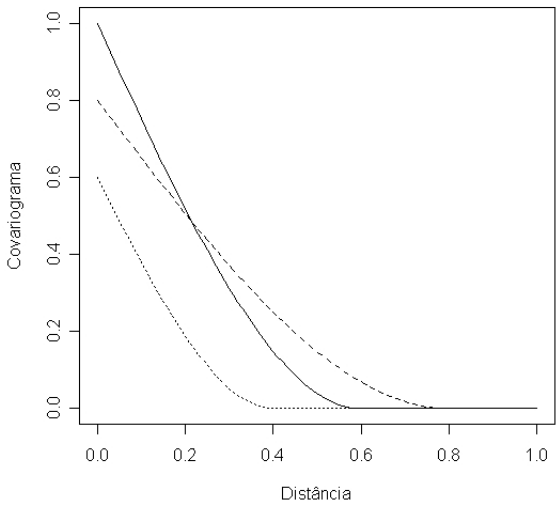

Figura 1.1: Representação gráfica de três covariogramas isotrópicos.

i) $\forall_{\mathbf{s}\in D} \quad C(\mathbf{0}) = \mathrm{Var}[Z(\mathbf{s})]$;

ii) $\forall_{\mathbf{h}\in\mathbb{R}^d} \quad C(\mathbf{h}) = C(-\mathbf{h})$;

iii) o covariograma é definido positivo, isto é,

$$\forall_{n\in\mathbb{N}}\ \forall_{a_1,\ldots,a_n\in\mathbb{R}}\ \forall_{\mathbf{s}_1,\ldots,\mathbf{s}_n\in D} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0. \tag{1.3.1}$$

A propriedade (1.3.1) resulta do facto de $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j)$ ser igual a $\mathrm{Var}\left[\sum_{i=1}^{n} a_i Z(\mathbf{s}_i)\right]$ que é sempre não negativa.

Frequentemente, para cada direcção fixa, existe um valor de distância a partir do qual o covariograma é nulo. Como se verá no seguimento, esse valor vai definir a amplitude do variograma.

Sendo $C(\mathbf{0})$ finito, a função covariograma pode ser substituída pela função $\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0})$, que se chama o correlograma. No entanto, a informação traduzida pela função correlograma é em tudo semelhante à da função covariograma. Contudo, como se verá a seguir, existem algumas vantagens em usar a função variograma em substituição do covariograma.

Considere-se agora a função variograma, que traduz a variância dos incrementos do processo.

Definição 1.3.2. Para qualquer localização $\mathbf{s} \in D$ e para qualquer vector $\mathbf{h} \in \mathbb{R}^d$ tal que $\mathbf{s} + \mathbf{h} \in D$, chama-se *variograma* à função

$$\begin{array}{rccl} 2\gamma\colon & \mathbb{R}^d & \longrightarrow & \mathbb{R}_0^+ \\ & \mathbf{h} & \longmapsto & 2\gamma(\mathbf{h}) = \mathrm{Var}[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})]. \end{array}$$

$\Diamond$

Por vezes utiliza-se o semivariograma em vez do variograma, uma vez que ambas as funções gozam das mesmas propriedades.

Repare-se que o variograma é definido para qualquer processo intrinsecamente estacionário, enquanto que o covariograma apenas existe em processos estacionários de segunda ordem. Portanto, sempre que um processo tem covariograma, então ele também tem variograma. Contudo, o recíproco não é verdadeiro. Esta é a grande vantagem que o variograma tem em relação ao covariograma, e é o principal motivo que leva à preferência pelo variograma como instrumento fundamental no desenvolvimento da Geoestatística.

Tal como para o covariograma, o variograma diz-se *isotrópico* se depender apenas da norma do vector $\mathbf{h}$. Nesse caso, o variograma não depende da direcção de $\mathbf{h}$ e, por isso, é o mesmo em todas as direcções.

Em grande parte dos casos, o variograma é uma função crescente, pelos motivos já referidos para justificar que o covariograma é frequentemente uma função decrescente. A Figura 1.2 mostra três exemplos de variogramas isotrópicos.

De seguida, apresentam-se as principais propriedades do variograma.

i) $2\gamma(\mathbf{0}) = 0$;

ii) $\forall_{\mathbf{h}\in\mathbb{R}^d} \quad 2\gamma(\mathbf{h}) = 2\gamma(-\mathbf{h})$;

iii) para qualquer $n \in \mathbb{N}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0, \tag{1.3.2}$$

onde $a_1, ..., a_n$ são tais que $\sum_{i=1}^{n} a_i = 0$ e $\mathbf{s}_1, ..., \mathbf{s}_n \in D$.

A restrição $\sum a_i = 0$ motiva a designação usual de que o variograma é *condicionalmente definido negativo*;

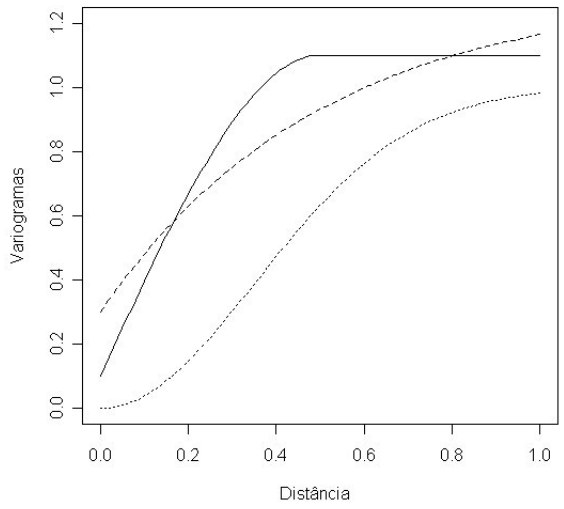

Figura 1.2: Representação gráfica de três variogramas isotrópicos.

iv) o variograma não pode crescer a um ritmo superior a $\|\mathbf{h}\|^2$, isto é,

$$\lim_{\|\mathbf{h}\|\to\infty} \frac{2\gamma(\mathbf{h})}{\|\mathbf{h}\|^2} = 0.$$

Esta condição aparece designada, na literatura, por *hipótese intrínseca*.

Para compreender a propriedade (1.3.2), note.se que a restrição $\sum_{i=1}^{n} a_i = 0$ implica que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2 = -2\left(\sum_{i=1}^{n} a_i Z(\mathbf{s}_i)\right)^2.$$

Calculando as esperanças em ambos os membros da equação anterior e tendo em conta a estacionaridade da média, obtém-se que

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) = -2\mathrm{Var}\left[\sum_{i=1}^{n} a_i Z(\mathbf{s}_i)\right] \leq 0.$$

Como se pode verificar em Schabenberger e Gotway (2005), a hipótese intrínseca implica que, em processos estacionários de segunda ordem, a segunda derivada do covariograma calculada em zero seja um número real negativo. Este facto é importante porque garante que $C(\mathbf{0})$ é um máximo da função covariograma.

Se um variograma satisfizer as propriedades (i) a (iv), ele diz-se um *variograma válido*. Qualquer função que seja um variograma válido pode ser utilizada como variograma de algum processo geoestatístico. Como se pode ver em Cressie (1993), se um variograma é válido em $\mathbb{R}^{d+1}$, então ele também é válido em $\mathbb{R}^d$ ($d \geq 1$). No entanto, a implicação recíproca não se verifica no caso geral.

Em geral, à medida que a distância entre observações aumenta, o valor do semivariograma tende para um valor constante (*vide* Figura 1.2). A esse valor chama-se o patamar. Tanto o patamar como a amplitude (já atrás referida) são estatisticamente interpretados como parâmetros do variograma. Seguidamente, passa-se a apresentar a definição e a interpretação dos parâmetros estatísticos do variograma.

Definição 1.3.3. Seja γ o semivariograma de um processo geoestatístico $Z(\mathbf{s})$. Chama-se *patamar* ao valor $\lim_{\|\mathbf{h}\|\to+\infty} \gamma(\mathbf{h})$, caso este limite exista e seja finito.

$\Diamond$

O patamar é uma característica exclusiva de variogramas de processos estacionários de segunda ordem, uma vez que em variogramas de processos apenas intrinsecamente estacionários, este parâmetro não existe. O mesmo acontece com a amplitude.

Definição 1.3.4. Chama-se *amplitude* ou *alcance* do variograma $2\gamma(\mathbf{h})$ na direcção do vector $\mathbf{h}_0$ ao menor valor de $\|\mathbf{h}\|$ que satisfaz a condição

$$\forall_{\varepsilon>0} \quad C\left((1+\varepsilon)\mathbf{h}\right) = 0, \tag{1.3.3}$$

onde $\mathbf{h} = k\mathbf{h}_0, k \in \mathbb{R}$. A amplitude será denotada por ϕ.

$\Diamond$

Quando existe amplitude na direcção de $\mathbf{s}_i - \mathbf{s}_j$, todas as variáveis aleatórias $Z(\mathbf{s}_i)$ e $Z(\mathbf{s}_j)$ para as quais a norma $\|\mathbf{s}_i - \mathbf{s}_j\|$ é superior à amplitude, são não correlacionadas. No entanto, a amplitude nem sempre existe. Por exemplo, existem modelos de covariogramas que só tomam o valor nulo assintoticamente. Nesses casos, não existe amplitude, mas pode sempre ser definida uma amplitude prática do variograma: a *amplitude prática* (também designada por *alcance efectivo*) é a distância a partir da qual a correlação que existe entre as variáveis aleatórias pode ser desprezada; por isso, as variáveis aleatórias que estão separadas por uma distância superior à amplitude

prática, podem ser consideradas não correlacionadas. Em Schabenberger e Gotway (2005), quando um processo com $C(\mathbf{0}) < \infty$ não tem amplitude na direcção de $\mathbf{h}_0$, define-se a amplitude prática como sendo o menor valor de $\|\mathbf{h}\|$ tal que, para qualquer ε positivo, $|C((1+\varepsilon)\mathbf{h})| \leq 0.05 \times C(\mathbf{0})$.

A condição (1.3.3) também pode ser expressa em termos do variograma. Assim, a amplitude pode ser definida como o menor valor de $\|\mathbf{h}\|$ que faz com que, para qualquer ε positivo, $\gamma((1+\varepsilon)\mathbf{h})$ seja igual ao patamar.

Consequentemente, o patamar e a amplitude são dois conceitos que estão interligados. Só pode existir amplitude quando existe patamar. Por isso, tal como para o patamar, só existe amplitude em variogramas de processos estacionários de segunda ordem.

O variograma nem sempre é uma função contínua. É mesmo frequente que o variograma tenha uma descontinuidade na origem. Esta descontinuidade pode ser originada pelo erro presente em cada medição, o que se traduz pela possibilidade de duas observações, na mesma localização, poderem ser bastante distintas; a descontinuidade também pode ser originada por uma variação em microescala do processo que o variograma não está a conseguir modelar. Surge assim um outro parâmetro do variograma, designado por efeito de pepita.

Definição 1.3.5. Seja γ o semivariograma de um processo geoestatístico $Z(\mathbf{s})$. Chama-se *efeito de pepita* ao valor $\tau^2 = \lim_{\|\mathbf{h}\|\to 0} \gamma(\mathbf{h})$.

$\Diamond$

A possibilidade de se considerar a existência do efeito de pepita é outra vantagem que o variograma apresenta em relação ao covariograma. Se o processo $Z(\mathbf{s})$ for estacionário de segunda ordem e o efeito de pepita não for nulo, a equação (1.2.5) tem que ser substituída por

$$\forall_{\mathbf{h}\in\mathbb{R}^d\setminus\{\mathbf{0}\}} \quad \gamma(\mathbf{h}) = \tau^2 + C(\mathbf{0}) - C(\mathbf{h}). \tag{1.3.4}$$

O comportamento do variograma junto da origem é fundamental para caracterizar a regularidade do processo $Z(\mathbf{s})$. Nesse sentido,

- se o variograma é contínuo na origem, *i.e.*, se $\tau^2 = 0$, então o processo $Z(\mathbf{s})$ é L_2-contínuo, ou seja, $\mathrm{E}\left[(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2\right] \to 0$ à medida que $\|\mathbf{h}\| \to 0$;

- se o efeito de pepita é positivo, ou seja, se o variograma é descontínuo na origem, o processo $Z(\mathbf{s})$ não é L_2-contínuo e, por isso, na vizinhança de cada ponto do domínio, o processo é altamente irregular;

- se o variograma for constante em todo o seu domínio excepto na origem, então, sempre que $\mathbf{s}_i \neq \mathbf{s}_j$, as variáveis aleatórias $Z(\mathbf{s}_i)$ e $Z(\mathbf{s}_j)$ não são correlacionadas.

Quando o variograma tem efeito de pepita positivo, então o processo $Z(\mathbf{s})$ pode ser decomposto na soma de vários subprocessos

$$\forall_{\mathbf{s}\in D}\ Z(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \eta(\mathbf{s}) + \epsilon(\mathbf{s}),$$

onde:

- $\mu(\mathbf{s}) = \mathrm{E}[Z(\mathbf{s})]$ é uma componente determinística, relativa à média de $Z(\mathbf{s})$, designada por variação em larga escala. Quando o processo $Z(\mathbf{s})$ é estacionário, esta componente é constante;

- $W(\mathbf{s})$ é um processo de média nula, L_2-contínuo e intrinsecamente estacionário. Se o variograma deste processo tiver amplitude, ela tem que ser superior ao $\min\{\|\mathbf{s}_i - \mathbf{s}_j\| : 1 \leq i < j \leq n\}$. Este processo é designado por variação suave em pequena escala;

- $\eta(\mathbf{s})$ é um processo de média nula, estacionário de segunda ordem e independente do processo $W(\mathbf{s})$. O variograma de $\eta(\mathbf{s})$ tem amplitude inferior ao $\min\{\|\mathbf{s}_i - \mathbf{s}_j\| : 1 \leq i < j \leq n\}$. Este processo é designado por variação em microescala e é o principal responsável pela existência de efeito de pepita positivo;

- $\epsilon(\mathbf{s})$ representa um processo de ruído branco, isto é, um processo de variáveis aleatórias não correlacionadas com média nula e variância constante. É um processo independente de $W(\mathbf{s})$ e de $\eta(\mathbf{s})$ que se designa por ruído ou por erro de medição. O ruído também pode ser causador de efeito de pepita positivo.

O efeito de pepita é originado pela soma da variância da variação em microescala com a variância do erro de medição, ou seja, $\tau^2 = \mathrm{Var}[\eta(\mathbf{s})] + \mathrm{Var}[\epsilon(\mathbf{s})]$.

A Figura 1.3 apresenta um semivariograma onde são salientados os parâmetros típicos dos processos estacionários de segunda ordem. Pode-se visualizar a amplitude

ϕ, o efeito de pepita τ^2 e o patamar $\tau^2 + \sigma^2$. O valor $\sigma^2 = \text{Var}[W(\mathbf{s})]$ é denominado patamar parcial uma vez que é igual ao patamar, sempre que o efeito de pepita é nulo.

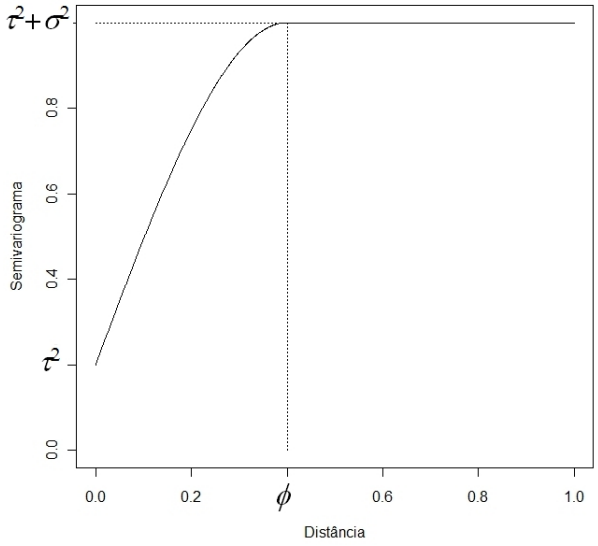

Figura 1.3: Ilustração dos parâmetros típicos dos semivariogramas de processos estacionários de segunda ordem.

Para finalizar esta secção, conclui-se que o variograma apresenta vantagens significativas em relação ao covariograma, nomeadamente, porque existe em mais processos geoestatísticos do que o covariograma e porque tem a possibilidade de incorporar o efeito de pepita, o qual permite que não se ignorem os efeitos da variação em micro-escala e do erro de medição. Por isso, ao longo deste trabalho vai-se estudar apenas o variograma.

1.3.2 Modelos de variograma isotrópico

Seguidamente, vão-se apresentar os modelos paramétricos mais conhecidos de entre os variogramas isotrópicos. Estes variogramas encontram-se agrupados por famílias. Cada família é constituída por um conjunto de variogramas definidos pela mesma expressão algébrica, a qual depende do parâmetro vectorial $\boldsymbol{\theta} \in \Theta$, onde Θ representa o espaço do parâmetro.

Modelo linear:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \theta_1 \|\mathbf{h}\|, & se \quad \|\mathbf{h}\| > 0 \end{cases},$$

onde $\boldsymbol{\theta} = (\tau^2, \theta_1)$ e $\theta_1 > 0$. Este modelo é válido para qualquer dimensão $d \geq 1$. No entanto, ele não possui patamar, o que torna impossível a sua utilização em processos estacionários de segunda ordem.

Modelo de tenda:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \frac{\sigma^2\|\mathbf{h}\|}{\phi}, & se \quad 0 < \|\mathbf{h}\| \leq \phi \\ \tau^2 + \sigma^2, & se \quad \|\mathbf{h}\| > \phi \end{cases}, \tag{1.3.5}$$

onde $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Este modelo é válido apenas em $\mathbb{R}$ e, por isso, é pouco utilizado em Geoestatística.

Modelo circular:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \sigma^2(1 - \rho(\|\mathbf{h}\|)), & se \quad 0 < \|\mathbf{h}\| \leq \phi \\ \tau^2 + \sigma^2, & se \quad \|\mathbf{h}\| > \phi \end{cases}, \tag{1.3.6}$$

onde $\rho(\|\mathbf{h}\|) = \frac{2}{\pi}\left(\arccos\left(\frac{\|\mathbf{h}\|}{\phi}\right) - \frac{\|\mathbf{h}\|}{\phi}\sqrt{1 - \frac{\|\mathbf{h}\|^2}{\phi^2}}\right)$, $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Este modelo só é válido em $\mathbb{R}$ e em $\mathbb{R}^2$, pelo que não pode ser utilizado em processos de domínio tridimensional.

Modelo esférico:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \sigma^2\left[\frac{3\|\mathbf{h}\|}{2\phi} - \frac{\|\mathbf{h}\|^3}{2\phi^3}\right], & se \quad 0 < \|\mathbf{h}\| \leq \phi \\ \tau^2 + \sigma^2, & se \quad \|\mathbf{h}\| > \phi \end{cases}, \tag{1.3.7}$$

onde $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Este modelo é válido apenas em $\mathbb{R}$, $\mathbb{R}^2$ e $\mathbb{R}^3$ mas, de acordo com Soares (2000), é um dos modelos mais utilizados nas aplicações da Geoestatística.

Modelo de Matérn:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \sigma^2\left[1 - \frac{(\|\mathbf{h}\|/\phi)^\kappa K_\kappa(\|\mathbf{h}\|/\phi)}{2^{\kappa-1}\Gamma(\kappa)}\right], & se \quad \|\mathbf{h}\| > 0 \end{cases},$$

onde $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi, \kappa)$ com $\phi > 0$ e $\kappa > 0$, K_κ é a função de Bessel de ordem κ e Γ é a função Gama. Este modelo é válido em qualquer dimensão $d \geq 1$. Repare-se que, ao contrário do que acontecia nos três modelos anteriores, neste modelo o parâmetro ϕ deixa de representar a amplitude e passa a representar a amplitude prática. O mesmo acontece nos modelos que se vão apresentar seguidamente.

Modelo exponencial:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \sigma^2\left[1 - e^{-3\frac{\|\mathbf{h}\|}{\phi}}\right], & se \quad \|\mathbf{h}\| > 0 \end{cases}, \tag{1.3.8}$$

onde $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Este modelo é um caso particular do modelo de Matérn para $\kappa = 0.5$.

Modelo Gaussiano:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0, & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \sigma^2\left[1 - e^{-3\frac{\|\mathbf{h}\|^2}{\phi^2}}\right], & se \quad \|\mathbf{h}\| > 0 \end{cases}, \tag{1.3.9}$$

onde $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Este modelo resulta do modelo de Matérn quando κ tende para $+\infty$.

Modelo de onda:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0 & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \sigma^2\left[1 - \frac{\phi}{\|\mathbf{h}\|}\sin\left(\frac{\|\mathbf{h}\|}{\phi}\right)\right] & se \quad \|\mathbf{h}\| > 0 \end{cases},$$

onde $\boldsymbol{\theta} = (\tau^2, \sigma^2, \phi)$ e $\phi > 0$. Este é um modelo especial uma vez que permite correlações negativas entre as variáveis aleatórias do processo. A correlação negativa tem interesse em processos com características de periodicidade. É um modelo válido apenas em $\mathbb{R}$, $\mathbb{R}^2$ e $\mathbb{R}^3$.

Modelo de potência:

$$\gamma(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} 0 & se \quad \|\mathbf{h}\| = 0 \\ \tau^2 + \theta\|\mathbf{h}\|^\lambda & se \quad \|\mathbf{h}\| > 0 \end{cases}, \tag{1.3.10}$$

onde $\boldsymbol{\theta} = (\tau^2, \theta, \lambda)$, $\theta > 0$ e $0 < \lambda < 2$. É um modelo que corresponde a variogramas de processos que são intrinsecamente estacionários, mas que

não são estacionários de segunda ordem. Por isso, os processos que são modelados por este variograma não apresentam covariograma. Este modelo é válido em qualquer dimensão e tem como caso particular o modelo linear, que se obtém tomando $\lambda = 1$.

Alguns exemplos dos modelos de semivariograma apresentados anteriormente podem ser visualizados na Figura 1.4.

Se o processo $Z(\mathbf{s})$ for estacionário de segunda ordem, é possível transformar um modelo válido de semivariograma num modelo válido de covariograma, através da equação (1.3.4). Como exemplo, se um processo estacionário de segunda ordem tiver um semivariograma esférico, então o seu covariograma será da forma

$$C(\|\mathbf{h}\|;\boldsymbol{\theta}) = \begin{cases} \sigma^2 \left[1 - \frac{3\|\mathbf{h}\|}{2\phi} + \frac{\|\mathbf{h}\|^3}{2\phi^3}\right], & se \quad 0 \leq \|\mathbf{h}\| < \phi \\ 0, & se \quad \|\mathbf{h}\| \geq \phi \end{cases}.$$

Em Journel e Huijbregts (1978) podem-se encontrar informações mais detalhadas sobre modelos paramétricos de variogramas.

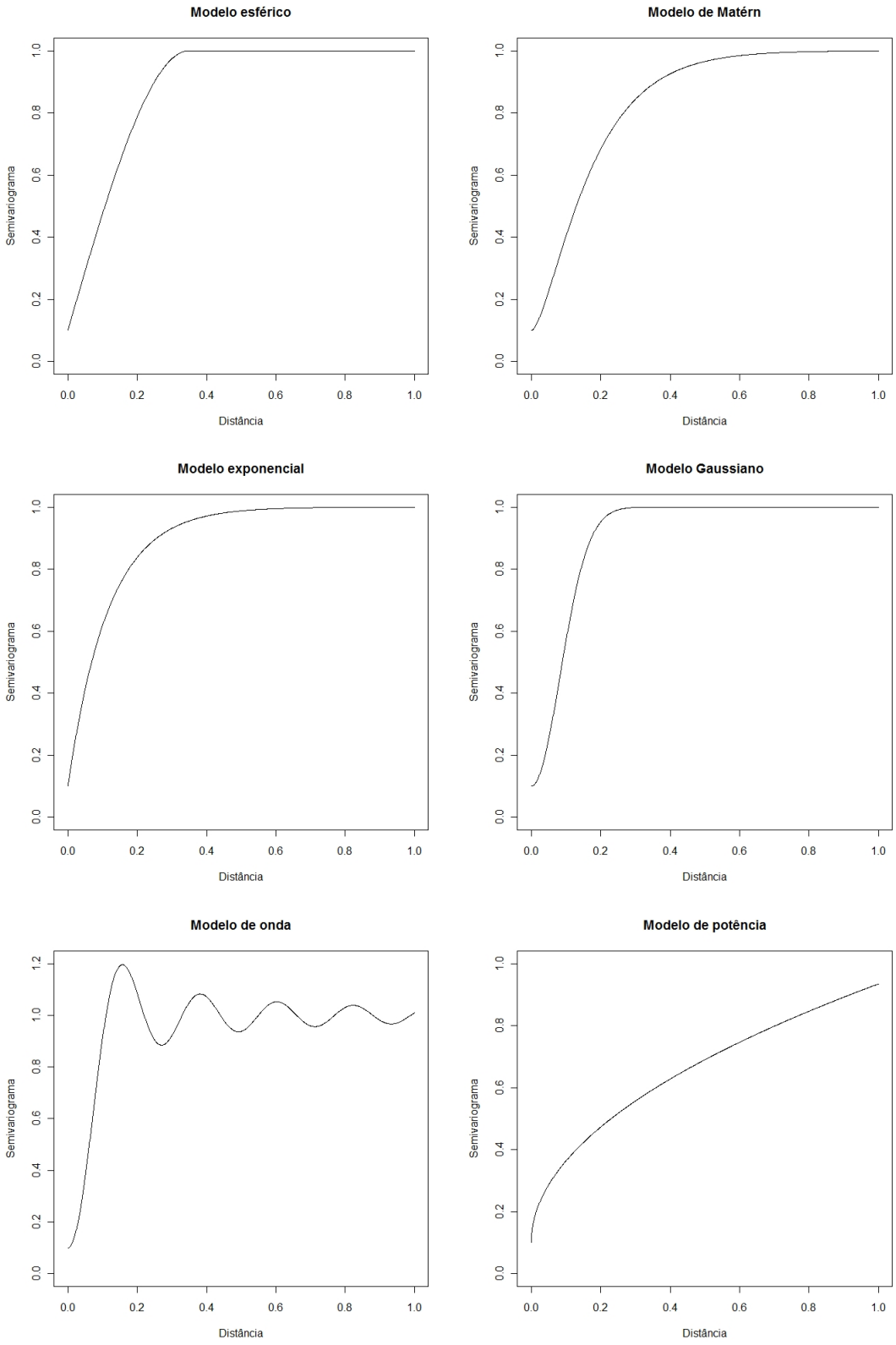

Figura 1.4: Ilustração das curvas de semivariograma para alguns modelos isotrópicos.

Capítulo 2

Métodos usuais na estimação da estrutura de dependência

Como se viu ao longo da secção **1.3**, se um processo geoestatístico $Z(\mathbf{s})$ for, pelo menos, intrinsecamente estacionário, a sua estrutura de dependência pode ser caracterizada através do variograma. Além disso, também foi realçada a importância do variograma como sendo o melhor instrumento para descrever a estrutura de dependência de um processo geoestatístico.

Assim, justifica-se que o presente capítulo seja inteiramente dedicado à descrição da metodologia tradicional de estimação do variograma.

O processo usual de estimação do variograma é constituído por duas etapas fundamentais que se podem resumir do seguinte modo:

1. a partir da amostra do processo, obtém-se um conjunto finito de estimativas pontuais da função variograma, as quais são determinadas em pontos específicos do domínio da função. Alguns autores chamam a este conjunto de estimativas pontuais do variograma, o variograma empírico, designação essa que por vezes também será utilizada no seguimento;

2. na segunda etapa, utiliza-se o variograma empírico para ajustar o modelo (teórico) de variograma mais adequado, procedendo à estimação dos parâmetros desse modelo.

Nas secções **2.1** e **2.2** serão tratadas cada uma das etapas do processo de estimação.

2.1 Obtenção de estimativas pontuais do variograma

Considere-se que, a partir de uma amostra $\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)\}$ do processo $Z(\mathbf{s})$, se pretende obter um conjunto finito de estimativas pontuais da função variograma. Tais estimativas pontuais do variograma são tradicionalmente obtidas através do estimador proposto por Matheron (1962), o qual tem expressão

$$2\hat{\gamma}(\mathbf{h}) = \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} \left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2, \tag{2.1.1}$$

onde $N(\mathbf{h})$ é o conjunto de pares de localizações cuja diferença é igual ao vector $\mathbf{h}$, *i.e.*,

$$N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}; i, j = 1, ..., n\}. \tag{2.1.2}$$

Note-se que o estimador de Matheron é obtido pelo método dos momentos, uma vez que resulta de igualar a variância dos incrementos à sua variância amostral.

Facilmente se verifica que o estimador $2\hat{\gamma}(\mathbf{h})$ é centrado em processos intrinsecamente estacionários, pois para qualquer vector $\mathbf{h}$ fixo,

$$\begin{aligned}
\mathrm{E}[2\hat{\gamma}(\mathbf{h})] &= \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} \mathrm{E}\left[\left(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\right)^2\right] \\
&= \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} \mathrm{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] \text{ (pela estacionaridade da média)} \\
&= \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \\
&= \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} 2\gamma(\mathbf{h}) \\
&= 2\gamma(\mathbf{h}).
\end{aligned}$$

Como é frequente que as amostras reais tenham localizações dispostas de uma forma irregular, é comum encontrarem-se situações onde, para um vector $\mathbf{h}$ fixo, o conjunto $N(\mathbf{h})$ tem poucos elementos. Isso faz com que seja difícil obter uma estimativa pontual precisa do variograma nesse vector $\mathbf{h}$. Para contornar essa dificuldade, na prática, agrupam-se os vectores $\mathbf{h}$ em regiões de tolerância. Deste modo, para estimar os pontos do variograma, em vez de se considerar $N(\mathbf{h})$, considera-se o conjunto $N(T(\mathbf{h}))$, onde

$T(\mathbf{h})$ é um conjunto de vectores que se encontram numa região de tolerância de $\mathbb{R}^d$ do vector $\mathbf{h}$. Expresso de outro modo,

$$N(T(\mathbf{h})) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \delta\mathbf{h}; \delta \in]1-\varepsilon, 1+\varepsilon[; i, j = 1, ..., n\},$$

para ε positivo. Repare-se que existem outras regiões de tolerância que também podem ser consideradas.

A Figura 2.1 representa o conjunto das estimativas pontuais de um semivariograma isotrópico. Neste caso, $N(T(\mathbf{h}))$ depende de $\mathbf{h}$ apenas através de $\|\mathbf{h}\|$ e as regiões de tolerância foram definidas por

$$N(T(\|\mathbf{h}\|)) = \{(\mathbf{s}_i, \mathbf{s}_j) : \|\mathbf{s}_i - \mathbf{s}_j\| \in]\,\|\mathbf{h}\| - 0.5, \|\mathbf{h}\| + 0.5[; i, j = 1, ..., n\}.$$

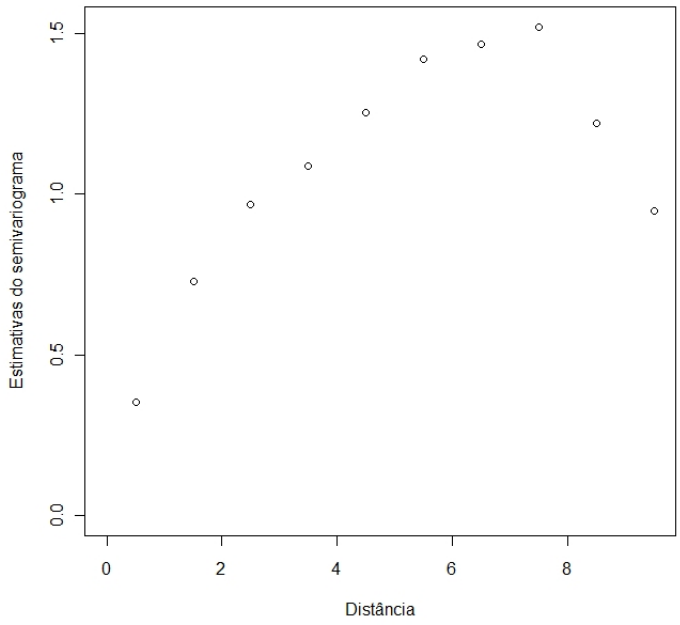

Figura 2.1: Representação das estimativas pontuais de um semivariograma isotrópico.

Apesar do agrupamento dos vectores em regiões de tolerância ser quase sempre efectuado na prática, este procedimento gera alguma ambiguidade na estimação do variograma. Repare-se que, para cada região de tolerância considerada, se obtém um conjunto diferente de estimativas pontuais do variograma. Por isso, devem-se escolher

regiões de tolerância apropriadas uma vez que, se a tolerância for pequena, corre-se o risco das estimativas pontuais serem calculadas a partir de poucos pontos, tornando-se assim pouco precisas; por outro lado, quanto maior for a tolerância, mais grosseiro se torna o conjunto de estimativas pontuais do variograma, uma vez que deixa de transparecer a estrutura de dependência que existe dentro das regiões de tolerância consideradas. Segundo Journel e Huijbregts (1978), para que as estimativas pontuais do variograma sejam precisas, devem ser estimadas com base em, pelo menos, 30 observações, ou seja, $\#N(T(\mathbf{h})) \geq 30$.

A ideia de agrupar os vectores por regiões de tolerância abriu a porta ao estudo de estimadores não paramétricos do variograma, os quais são obtidos através do método do núcleo. Nestes estimadores, todos os incrementos $Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$ são utilizados para estimar o variograma num determinado vector $\mathbf{h}$. Cada incremento tem associado um determinado peso, que é tanto maior quanto mais próximo for o vector $\mathbf{s}_i - \mathbf{s}_j$ do vector $\mathbf{h}$. Este procedimento também é conhecido como o método de Nadaraya-Watson. A vantagem destes estimadores é que eles podem estimar o variograma em qualquer vector $\mathbf{h}$ considerado. Contudo, normalmente eles não são centrados, contrariamente ao estimador de Matheron definido em (2.1.1). Repare-se que o estimador de Matheron com regiões de tolerância, também pode ser encarado como um estimador não paramétrico obtido pelo método do núcleo – neste caso, os pesos atribuídos aos incrementos só são 1 ou 0, consoante o incremento entra ou não para determinar a estimativa.

Infelizmente, nem os estimadores não paramétricos obtidos através do método do núcleo, nem o estimador de Matheron, dão a garantia de devolver estimativas de acordo com um variograma válido. De facto, o variograma tem que ser condicionalmente definido negativo, o que nem sempre se verifica com os variogramas empíricos obtidos com estes estimadores. A solução para este problema, passa por encontrar uma função variograma válida que se aproxime o mais possível das estimativas obtidas. Como o estimador de Matheron é centrado e os estimadores obtidos pelo método do núcleo, geralmente, só são assintoticamente centrados, é o estimador de Matheron que mais se utiliza para obter as estimativas pontuais do variograma.

Depois de encontradas as estimativas pontuais do variograma, entra-se na segunda etapa de estimação do variograma.

2.2 Estimação dos parâmetros de um modelo de variograma

O conjunto de estimativas pontuais do variograma, cuja obtenção foi descrita na secção anterior, não é suficiente para traduzir a estrutura de dependência de um processo geoestatístico, uma vez que não tem domínio em $\mathbb{R}^d$, e que nem sempre é condicionalmente definido negativo. Assim, para se obter uma estimativa válida do variograma, é necessário ajustar um modelo paramétrico que seja adequado às estimativas pontuais encontradas.

A escolha do modelo de variograma que melhor se ajusta à estrutura de dependência revelada pelas observações do processo, não é uma tarefa fácil, uma vez que o conjunto de variogramas válidos é bastante vasto. O procedimento mais utilizado na prática, consiste em escolher, com base em conhecimentos empíricos, uma família de modelos do conjunto de todos os variogramas paramétricos válidos

$$\{2\gamma : 2\gamma(\mathbf{h}) = 2\gamma(\mathbf{h};\boldsymbol{\theta}); \mathbf{h} \in \mathbb{R}^d, \boldsymbol{\theta} \in \Theta\}.$$

Actualmente existem alguns métodos estatísticos que podem ajudar o investigador no processo de selecção do modelo de variograma. Como exemplo, Gorsich e Genton (2000) propuseram uma metodologia de escolha do modelo de variograma com base na comparação de derivadas. Na opinião desses autores, os modelos de variograma diferem muito pouco entre si – o que difere, significativamente, é a derivada de cada uma das funções 2γ, de acordo com os modelos. Consequentemente, esses autores defendem que o modelo de variograma que é mais adequado ao processo em estudo, é aquele cuja derivada se aproxima mais da estimativa da derivada do modelo, a qual se obtém usando a estimativa não paramétrica encontrada na primeira etapa de estimação. Por outro lado, Maglione e Diblasi (2001) desenvolveram um teste para decidir se um determinado modelo de variograma é adequado para modelar um dado processo em estudo.

Depois de se ter escolhido uma família de variogramas, estimam-se os parâmetros de modo a que a função variograma se aproxime o mais possível das estimativas pontuais do variograma, as quais foram obtidas na primeira etapa do processo de estimação.

Por outras palavras, determina-se $\hat{\boldsymbol{\theta}} \in \Theta$ que faz com que $2\gamma(\mathbf{h}; \hat{\boldsymbol{\theta}})$ optimize o critério de ajustamento escolhido. Existem diversos critérios de ajustamento, com base nos quais se encontram os estimadores mais utilizados, de entre os quais se salientam a máxima verosimilhança e os mínimos quadrados. Seguidamente apresentam-se os correspondentes métodos de estimação.

2.2.1 Método da máxima verosimilhança

O estimador de máxima verosimilhança é obtido, em geral, supondo a distribuição normal de $Z(\mathbf{s})$. Admitindo a estacionaridade da média e que as observações são provenientes de uma distribuição normal multivariada, com matriz de covariâncias $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, *i.e.*, que $\mathbf{Z} = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, então o logaritmo da função de verosimilhança vem expresso por

$$\log L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det(\boldsymbol{\Sigma}(\boldsymbol{\theta}))) + \frac{1}{2}(\mathbf{Z} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \boldsymbol{\mu}),$$

onde $\det(\boldsymbol{\Sigma}(\boldsymbol{\theta}))$ representa o determinante da matriz $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. O estimador de máxima verosimilhança, $\hat{\boldsymbol{\theta}}_{ML}$, obtém-se determinando o valor $\boldsymbol{\theta} \in \Theta$ que maximiza a função $\log L(\boldsymbol{\theta})$.

Repare-se que as estimativas pontuais do variograma não entram no processo de maximização da função de verosimilhança. De facto, de acordo com este método, tais estimativas apenas contribuem para seleccionar o modelo de variograma mais adequado.

Embora os estimadores de máxima verosimilhança tenham boas propriedades em geral, também têm algumas desvantagens. Para além de ser necessário pressupor o conhecimento da distribuição do processo, verifica-se que $\hat{\boldsymbol{\theta}}_{ML}$ é um estimador enviesado, como se pode ver em Cressie (1993).

Por estes motivos, o método da máxima verosimilhança é menos utilizado do que, por exemplo, o método dos mínimos quadrados e, por isso, não será considerado em detalhe no desenvolvimento deste trabalho.

2.2.2 Método dos mínimos quadrados

O método de mínimos quadrados é intuitivo, está muito divulgado e impõe hipóteses pouco restritivas sobre as distribuições das observações do processo. Por isso, não é de

admirar que seja um dos métodos mais utilizados nas aplicações da Geoestatística.

Considere-se um vector constituído pelas estimativas pontuais do variograma, $2\hat{\boldsymbol{\gamma}} = (2\hat{\gamma}(\mathbf{h}_1), ..., 2\hat{\gamma}(\mathbf{h}_H))$, para $H \in \mathbb{N}$, e um outro vector cujas componentes são definidas por um modelo de variograma paramétrico válido, $2\boldsymbol{\gamma}(\boldsymbol{\theta}) = (2\gamma(\mathbf{h}_1; \boldsymbol{\theta}), ..., 2\gamma(\mathbf{h}_H; \boldsymbol{\theta}))$, os quais foram obtidos nos mesmos pontos $\mathbf{h}_1, ..., \mathbf{h}_H$. O estimador de mínimos quadrados $\hat{\boldsymbol{\theta}}_{LS}$ é determinado pela solução $\boldsymbol{\theta} \in \Theta$ que minimiza uma expressão do tipo

$$(2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta}))^{\mathrm{T}} \mathbf{V}^{-1} (2\hat{\boldsymbol{\gamma}} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})), \tag{2.2.1}$$

onde $\mathbf{V}$ representa a matriz de covariâncias do estimador.

Se $\mathbf{V}$ for a matriz identidade de ordem H, então $\hat{\boldsymbol{\theta}}_{LS}$ corresponde ao estimador de mínimos quadrados simples (denotado por *OLS*); se $\mathbf{V}$ for uma matriz diagonal tal que $v_i = \mathrm{Var}[2\hat{\gamma}(\mathbf{h}_i)]$, o critério passa a ser designado por mínimos quadrados ponderados (denotado por *WLS*); finalmente, se $\mathbf{V}$ for uma matriz quadrada tal que $v_{i,j} = \mathrm{Cov}[2\hat{\gamma}(\mathbf{h}_i), 2\hat{\gamma}(\mathbf{h}_j)]$, o estimador resultante é referido como estimador de mínimos quadrados generalizados (denotado por *GLS*).

Tendo em conta as características do processo, a matriz $\mathbf{V}$ não é múltipla da identidade. Por isso, o estimador de *OLS* é desaconselhado, uma vez que não goza das propriedades que o tornam atractivo num cenário de observações *i.i.d.*. Por outro lado, não é possível conhecer a matriz $\mathbf{V}$, pois ela depende do próprio variograma que se pretende estimar.

No exemplo que se segue, particulariza-se o cálculo de $\mathbf{V}$ para o caso do estimador de Matheron, supondo um processo Gaussiano.

Exemplo 2.2.1. Seja $Z(\mathbf{s})$ um processo Gaussiano. A estacionaridade da média garante que

$$\forall_{\mathbf{s}\in D} \forall_{\mathbf{h}\in R^d} \; \frac{Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})}{\sqrt{2\gamma(\mathbf{h})}} \sim N(0,1),$$

e, portanto, que

$$\forall_{\mathbf{s}\in D} \forall_{\mathbf{h}\in R^d} \; \frac{(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s}))^2}{2\gamma(\mathbf{h})} \sim \chi^2_1, \tag{2.2.2}$$

onde χ^2_1 representa a distribuição do Qui-quadrado com um grau de liberdade.

Para facilitar a notação, considere-se que $\mathbf{h}_{i,j} = \mathbf{s}_i - \mathbf{s}_j$ e represente-se por $T_{i,j} = Z(\mathbf{s}_i) - Z(\mathbf{s}_j)$ o incremento do processo entre $\mathbf{s}_i$ e $\mathbf{s}_j$. Note-se, desde já, que por (2.2.2) se sabe que $T^2_{i,j}$ tem valor esperado $2\gamma(\mathbf{h}_{i,j})$ e variância $8\gamma^2(\mathbf{h}_{i,j})$.

Calculando a variância do estimador de Matheron apresentado em (2.1.1), para um qualquer vector $\mathbf{h}_m$ tem-se que

$$\begin{array}{rcl} \mathrm{Var}[2\hat{\gamma}(\mathbf{h}_m)] & = & \frac{1}{(\#N(\mathbf{h}_m))^2}\mathrm{Var}\left[\sum_{N(\mathbf{h}_m)} T_{i,j}^2\right] \\ & = & \frac{1}{(\#N(\mathbf{h}_m))^2}\sum_{i,j}\sum_{k,l}\mathrm{Cov}[T_{i.j}^2, T_{k.l}^2]. \end{array} \tag{2.2.3}$$

Contudo, é possível verificar que as parcelas $\mathrm{Cov}[T_{i.j}^2, T_{k.l}^2]$ podem ser expressas em termos da função 2γ. Para tal é preciso ter em conta um resultado simples, o qual pode ser encontrado, por exemplo, em Casella e Berger (2002). Esse resultado afirma que, se $X \sim N(\mu_X, \sigma_X^2), Y \sim N(\mu_Y, \sigma_Y^2), \mathrm{Cov}[X,Y] = \sigma_{X,Y}, \mathrm{Var}[X^2] = \sigma_{X^2}^2$ e $\mathrm{Var}[Y^2] = \sigma_{Y^2}^2$, então

$$\frac{\mathrm{Cov}[X^2, Y^2]}{\sigma_{X^2}\sigma_{Y^2}} = \frac{\sigma_{X,Y}^2}{\sigma_X^2\sigma_Y^2}.$$

Sendo assim, tomando $X = T_{i,j}$ e $Y = T_{k,l}$, facilmente se conclui que

$$\begin{array}{rl} \mathrm{Cov}[T_{i,j}^2, T_{k,l}^2] & = \sqrt{\mathrm{Var}\left[T_{i,j}^2\right]\mathrm{Var}\left[T_{k,l}^2\right]} \times \frac{\mathrm{Cov}[T_{i,j},T_{k,l}]^2}{\mathrm{Var}[T_{i,j}]\mathrm{Var}[T_{k,l}]} \\ & = \frac{\sqrt{8\gamma^2(\mathbf{h}_{i,j})8\gamma^2(\mathbf{h}_{k,l})}\mathrm{Cov}[T_{i,j},T_{k,l}]^2}{2\gamma(\mathbf{h}_{i,j})2\gamma(\mathbf{h}_{k,l})} \\ & = 2\mathrm{Cov}[T_{i,j}, T_{k,l}]^2 \\ & = 2[\mathrm{Cov}[Z(\mathbf{s}_i), Z(\mathbf{s}_k)] + \mathrm{Cov}[Z(\mathbf{s}_j), Z(\mathbf{s}_l)] - \mathrm{Cov}[Z(\mathbf{s}_i), Z(\mathbf{s}_l)] - \mathrm{Cov}[Z(\mathbf{s}_j), Z(\mathbf{s}_k)]]^2 \\ & = 2[\gamma(\mathbf{h}_{i,k}) + \gamma(\mathbf{h}_{j,l}) - \gamma(\mathbf{h}_{i,l}) - \gamma(\mathbf{h}_{j,k})]^2. \end{array}$$

Isto implica que as variâncias e covariâncias do estimador de Matheron dependem do próprio variograma. Então, a matriz $\mathbf{V}$ é dependente do próprio parâmetro $\boldsymbol{\theta}$ que se pretende estimar. Para além disso, e como se pôde observar, mesmo para o estimador de Matheron num processo Gaussiano, a forma da matriz $\mathbf{V}$ não é simples. Mais, a inversão de $\mathbf{V}$ e a minimização por *GLS* é, frequentemente, computacionalmente impossível (*vide* Lahiri *et al.* (2002)).

$\Diamond$

Tendo em vista ultrapassar a dificuldade de cálculo da matriz $\mathbf{V}$, Cressie (1993) recomenda uma alternativa. A ideia é considerar que o estimador de *WLS* é um bom compromisso entre o *OLS* e o *GLS* no que diz respeito à eficiência/facilidade de cálculo. O autor sugere que os elementos da diagonal principal da matriz $\mathbf{V}$ sejam aproximados por

$$v_m = \mathrm{Var}[2\hat{\gamma}(\mathbf{h}_m)] \approx \frac{2[2\gamma(\mathbf{h}_m;\boldsymbol{\theta})]^2}{\#N(\mathbf{h}_m)}. \tag{2.2.4}$$

Esta aproximação resulta de (2.2.3), assumindo que os $T^2_{i,j}$ têm entre si uma covariância nula.

Apesar de Cressie defender o uso do *WLS*, Zimmerman e Zimmerman (1991) desenvolveram estudos empíricos que indicam que o *OLS* e o *WLS* têm desempenhos semelhantes, pelo que consideram desnecessária a utilização do *WLS*.

A Figura 2.2 representa um conjunto de curvas de acordo com um modelo esférico, as quais foram obtidas com as estimativas pontuais do semivariograma da Figura 2.1. Cada curva foi determinada através de um método de estimação diferente. A linha a cheio representa a estimativa obtida por *OLS*, a curva a tracejado foi obtida utilizando *WLS*, com a aproximação de Cressie (1993) e, finalmente, a curva a ponteado foi obtida pelo método da máxima verosimilhança.

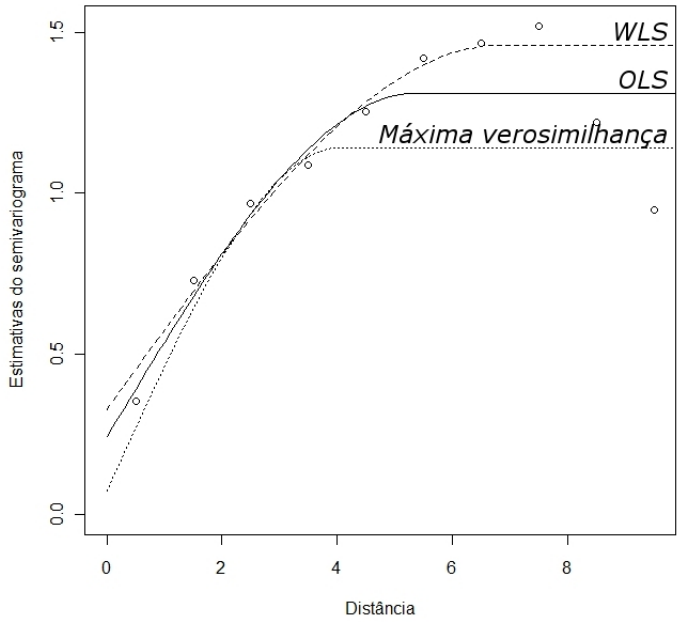

Figura 2.2: Representação das estimativas pontuais do semivariograma obtidas através do estimador de Matheron e dos semivariogramas do modelo esférico estimados por *WLS*, *OLS* e máxima verosimilhança.

É de salientar que, tal como verificaram Zimmerman e Zimmerman, as curvas obtidas por *OLS* e por *WLS* são bastante semelhantes, principalmente para valores pequenos de $\|\mathbf{h}\|$.

2.3 Propriedades assintóticas dos estimadores do variograma

Depois de apresentados os estimadores tradicionais do variograma, é essencial falar um pouco das suas propriedades assintóticas. Ao contrário do que acontece com outros modelos estatísticos, as características do domínio do processo geoestatístico – espacial e contínuo – fazem com que a dimensão da amostra possa aumentar de diferentes formas. Assim, existem diferentes metodologias de estudo, consoante o modo como se interpreta o aumento da dimensão da amostra.

2.3.1 Metodologias de estudo

O facto dos processos geoestatísticos terem índice num domínio espacial e contínuo permite que o número de observações da amostra tenda para infinito de três modos distintos:

i considerando mais observações dentro do próprio domínio inicial, isto é, que a distância entre as observações vai diminuindo e a densidade de observações vai aumentando (abordagem *infill asymptotics – IA*);

ii estendendo a $\mathbb{R}^d$ o domínio inicial do processo, mantendo a distância e a densidade de observações inalteradas (abordagem *increasing domain asymptotics – IDA*);

iii considerando uma mistura entre as duas últimas metodologias, isto é, permitindo que o domínio se estenda a $\mathbb{R}^d$, à medida que a densidade de observações vai aumentando (abordagem *mixed increasing domain asymptotics – MIDA*).

A ideia fundamental da *IA* resume-se a considerar que o domínio do processo $Z(\mathbf{s})$ é limitado e tem fronteiras fixas. Consequentemente, a única forma de obter resultados assintóticos é aumentando o número de observações dentro desse mesmo domínio. Cressie (1993) refere que este procedimento é muito utilizado, por exemplo, na indústria mineira, uma vez que, nestes casos, o terreno/domínio é fixo e limitado, e novas observações que se considerem através da exploração da mina, só podem estar localizadas

no mesmo terreno/domínio onde a mina está localizada. Então, a densidade das observações vai aumentando dentro do terreno considerado e, à medida que se vão supondo novas observações, elas serão cada vez mais próximas entre si.

Para formalizar a *IA* no caso em que as componentes da amostra estão dispostas ao longo de grelhas regulares, considera-se um conjunto

$$\mathfrak{X}^d = \{(\delta_1 i_1, ..., \delta_d i_d) : \delta_j > 0 \wedge i_j \in \mathbb{Z}\}, \quad j = 1, ..., d,$$

de localizações que estão dispostas ao longo de uma grelha d-dimensional, com incrementos δ_j na direcção j.

Sendo $D \subset \mathbb{R}^d$ o domínio de $Z(\mathbf{s})$ e $\{h_n\}_{n\in\mathbb{N}}$ uma sequência decrescente de números reais que tende para zero quando n tende para infinito, as localizações da n-ésima amostra são formalizadas por

$$\{\mathbf{s}_1, ..., \mathbf{s}_n\} = \{\mathbf{s} : \mathbf{s} \in h_n\mathfrak{X}^d \cap D\}.$$

À medida que n aumenta, a grelha $h_n\mathfrak{X}^d$ torna-se cada vez mais fina e a região amostrada mantém-se a mesma.

Por outro lado, a *IDA* segue uma abordagem análoga à que usualmente se utiliza nas séries temporais. Assim, a dimensão da amostra aumenta à custa do aumento da região que é considerada. Ou seja, de amostra para amostra, o domínio em estudo vai ficando cada vez maior, mas a distância e a densidade das observações mantêm-se inalteradas. Como refere Cressie (1993), esta metodologia assintótica pode ser utilizada, por exemplo, quando se está a estudar a produção das árvores de um pomar, mantendo o distanciamento entre árvores já existente. Como não se podem plantar mais árvores entre as árvores que existem inicialmente no pomar, a única hipótese de estudar o processo consiste em acrescentar árvores fora do terreno/domínio considerado inicialmente, seguindo o mesmo padrão das localizações já existentes.

Em Lahiri *et al.* (2002) pode-se encontrar a formalização da *IDA*. Para tal considera-se:

- uma região R_0 que é um subconjunto aberto de $]-1/2, 1/2]^d$ contendo a origem;
- uma sucessão $\{\lambda_n\}_{n\in\mathbb{N}}$ de números reais positivos que tende para infinito quando n tende para infinito;

- a sucessão de regiões $R_n = \lambda_n R_0, n \in \mathbb{N}$.

Tal como se pode ver no artigo referido, a região R_0 é como um protótipo das regiões R_n, as quais são obtidas ampliando λ_n vezes a região R_0. O facto da origem pertencer a R_0, permite que a forma das sucessivas regiões R_n se mantenha igual à de R_0. As localizações da n-ésima amostra são dadas por

$$\{\mathbf{s}_1, ..., \mathbf{s}_n\} = \{\mathbf{s} : \mathbf{s} \in \mathfrak{X}^d \cap R_n\}.$$

Neste caso, à medida que n aumenta, a grelha $\mathfrak{X}^d$ não se altera. A única coisa que se modifica é a região amostrada, a qual vai ficando cada vez maior.

Por fim, a *MIDA* é uma combinação das duas abordagens assintóticas apresentadas anteriormente.

A formalização deste procedimento é uma consequência imediata das formalizações anteriores. Deste modo, a n-ésima amostra é formada pelas localizações

$$\{\mathbf{s}_1, ..., \mathbf{s}_n\} = \{\mathbf{s} : \mathbf{s} \in h_n\mathfrak{X}^d \cap R_n\}.$$

Como se pode observar, à medida que n aumenta, a grelha $h_n\mathfrak{X}^d$ torna-se cada vez mais fina, enquanto a região R_n amostrada se torna cada vez maior.

Depois de apresentadas as diferentes abordagens, surge a questão de qual será a melhor. A resposta depende das características do processo em estudo, pelo que cabe ao investigador determinar qual é a metodologia assintótica mais apropriada em cada caso.

2.3.2 Consistência e distribuição assintótica dos estimadores

O procedimento tradicional de estimação do variograma recorre ao estimador de Matheron $(2\hat{\gamma}(\mathbf{h}_1), ..., 2\hat{\gamma}(\mathbf{h}_H))$, que foi apresentado em (2.1.1), para obter as estimativas pontuais do variograma. Como se verificou na secção **2.1**, o estimador é centrado para estimar o verdadeiro variograma $2\gamma(\mathbf{h}; \boldsymbol{\theta}_0)$ nos pontos de abcissa $\mathbf{h}_1, ..., \mathbf{h}_H$.

Para abordar a consistência do estimador de Matheron, considere-se que o processo $Z(\mathbf{s})$ é pelo menos intrinsecamente estacionário, tem observações normalmente

distribuídas e que os seus incrementos, $Z(\mathbf{s})-Z(\mathbf{s}+\mathbf{h})$, não são correlacionados. Nestas condições, como se referiu na subsecção **2.2.2**, tem-se que

$$\mathrm{Var}[2\hat{\gamma}(\mathbf{h}_i)] = 2\frac{(2\gamma(\mathbf{h}_i;\boldsymbol{\theta}_0))^2}{\#N(\mathbf{h}_i)}.$$

Deste modo, para qualquer uma das abordagens assintóticas apresentadas na subsecção anterior, à medida que a dimensão da amostra aumenta, $\#N(\mathbf{h}_i)$ tende para infinito e, consequentemente, $\mathrm{Var}[2\hat{\gamma}(\mathbf{h}_i)]$ tende para zero. Assim, uma vez que $2\hat{\gamma}(\mathbf{h}_i)$ é centrado e que, à medida que a dimensão da amostra aumenta, $\mathrm{Var}[2\hat{\gamma}(\mathbf{h}_i)]$ tende para zero, o estimador de Matheron converge em probabilidade para o variograma $2\gamma(\mathbf{h};\boldsymbol{\theta}_0)$ nos vectores $\mathbf{h}_i$ considerados.

Este procedimento para demonstrar a consistência do estimador de Matheron é simples, mas supõe que os incrementos do processo em estudo são não correlacionados. Contudo, os incrementos dos processos geoestatísticos são quase sempre correlacionados, o que torna esta hipótese bastante restritiva. Por isso, existem alternativas para demonstrar a consistência do estimador de Matheron, as quais consideram a correlação entre os incrementos.

Lahiri *et al.* (2002) chegam à consistência do estimador de Matheron através da sua distribuição assintótica. Eles mostraram que, quer por *IDA*, quer por *MIDA*, sob certas condições de regularidade fracas (onde não se inclui a não correlação dos incrementos do processo), o estimador de Matheron converge em lei para uma distribuição normal, isto é,

$$k_n\left(2\hat{\gamma}(\mathbf{h}_1) - 2\gamma(\mathbf{h}_1;\boldsymbol{\theta}_0), ..., 2\hat{\gamma}(\mathbf{h}_H) - 2\gamma(\mathbf{h}_H;\boldsymbol{\theta}_0)\right) \xrightarrow{\mathcal{L}} N_H\left(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\right),$$

onde a sequência $\{k_n\}_{n\in\mathbb{N}}$ e a matriz $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ dependem do contexto assintótico que se está a utilizar. Como, em ambos os contextos assintóticos considerados, a sequência $\{k_n\}_{n\in\mathbb{N}}$ tende para infinito, então o estimador de Matheron tende em probabilidade para o variograma $2\gamma(\mathbf{h};\boldsymbol{\theta}_0)$ nos vectores $\mathbf{h}_i$ considerados.

As considerações anteriores dizem respeito à primeira etapa da estimação. Como se referiu na secção **2.2**, a segunda etapa de estimação do variograma consiste na estimação do parâmetro $\boldsymbol{\theta} \in \Theta$ que faz com que um determinado modelo de variograma $2\gamma(\mathbf{h};\boldsymbol{\theta})$ se aproxime o mais possível das estimativas pontuais do variograma, obtidas

na primeira etapa. A segunda fase da estimação de $\boldsymbol{\theta}$ é tradicionalmente efectuada através do método de mínimos quadrados (já apresentados na subsecção **2.2.2**), o qual parece depender essencialmente da estrutura da matriz $\mathbf{V}$ das covariâncias dos erros do modelo. No entanto, surpreendentemente, a matriz $\mathbf{V}$ não é muito relevante para garantir as propriedades assintóticas dos estimadores de mínimos quadrados. Tal como se pode ver em Lahiri *et al.* (2002), as propriedades assintóticas dos estimadores de mínimos quadrados dependem, em grande parte, das propriedades do estimador pontual do variograma utilizado na primeira etapa. Nesse artigo, os autores mostraram que, sob condições de regularidade, se o estimador pontual do variograma converge para o verdadeiro variograma (em probabilidade ou quase certamente), então o estimador de mínimos quadrados também converge para os parâmetros do verdadeiro variograma (em probabilidade ou quase certamente), isto é,

$$2\hat{\gamma}(\mathbf{h}_i) - 2\gamma(\mathbf{h}_i;\theta_0) \xrightarrow{P} 0 \quad \Rightarrow \quad \hat{\theta}_{n,\mathbf{V}} - \theta_0 \xrightarrow{P} 0$$

e

$$2\hat{\gamma}(\mathbf{h}_i) - 2\gamma(\mathbf{h}_i;\theta_0) \xrightarrow{q.c.} 0 \quad \Rightarrow \quad \hat{\theta}_{n,\mathbf{V}} - \theta_0 \xrightarrow{q.c.} 0.$$

De modo análogo, a distribuição assintótica do estimador de mínimos quadrados também depende da distribuição assintótica do estimador pontual do variograma. No mesmo artigo e sob certas condições de regularidade, verifica-se que, se o estimador pontual do variograma segue assintoticamente uma distribuição normal, então o estimador de mínimos quadrados também segue a mesma distribuição assintótica.

Como consequência dos resultados dos parágrafos anteriores, dado que o estimador de Matheron é consistente e que segue assintoticamente uma distribuição normal, então os estimadores de mínimos quadrados também vão ser consistentes e também vão seguir assintoticamente uma distribuição normal.

Caso se prefira estimar o variograma através do método da máxima verosimilhança, também existem resultados conhecidos: sob um contexto de *IDA*, Mardia e Marshall (1984) mostraram que, se o processo $Z(\mathbf{s})$ tiver observações normais, o estimador de máxima verosimilhança é consistente e segue uma distribuição assintótica normal.

Capítulo 3

Robustez estatística

Este capítulo é dedicado à apresentação de noções fundamentais em estatística robusta, de modo a conter o material necessário para o desenvolvimento de estimadores robustos em modelos geoestatísticos. De seguida, a exposição centra-se na estimação robusta. Este tópico foi direccionado para o estudo de estimadores específicos, nomeadamente dos estimadores-MM e do estimador Q_n.

3.1 Introdução

Quando se utilizam modelos estatísticos, existe um conjunto de hipóteses ideais que são previamente assumidas como verdadeiras. Contudo, há muitas situações onde a realidade não se comporta de acordo com os modelos matemáticos, por isso, as hipóteses assumidas não se verificam, ou apenas se verificam aproximadamente. Quando isso acontece, os métodos estatísticos podem perder significativamente as boas propriedades que justificavam a sua utilização sob as condições supostas.

Geralmente, existe a ideia intuitiva de que, se a realidade não se afastar muito das hipóteses assumidas no modelo, as propriedades dos procedimentos não vão ser muito penalizadas. No entanto, esta ideia não está correcta, uma vez que está provado que pequenos afastamentos das hipóteses, podem conduzir a maus resultados.

Uma das situações em que existe afastamento das hipóteses do modelo e que ocorre com frequência nas aplicações, é a presença de *observações atípicas* na amostra. As observações atípicas, também designadas por *outliers*, são observações que se afastam do padrão revelado pelos restantes dados, tornando-se assim observações discordantes.

É necessário ter um cuidado especial com estas observações, porque elas podem indicar que o modelo não foi bem escolhido, ou podem mesmo ser o resultado de algum erro grosseiro. Por outro lado, a presença de observações atípicas na amostra, pode fazer com que a maior parte dos procedimentos tradicionais não conduzam a resultados desejáveis, mas torna-se difícil decidir o que fazer com essas observações.

Para além disso, hoje em dia, cada vez mais se exige que os métodos estatísticos sejam capazes de modelar uma grande quantidade de variáveis aleatórias, partindo de um conjunto reduzido de observações. Nestes casos, a detecção de *outliers* torna-se muito complicada, mesmo quando se utilizam as técnicas existentes para esse efeito. Repare-se que as técnicas de detecção de *outliers* têm algumas desvantagens. Elas dependem do modelo assumido (o qual pode não se verificar); podem sofrer do chamado *masking effect*, isto é, a detecção de um *outlier* faz com que outros não sejam detectados; e podem revelar o chamado *swamping effect*, ou seja, a detecção de um outlier tende a arrastar consigo outras observações. Por isso, as técnicas de detecção e posterior rejeição de *outliers*, seguidas de um método tradicional, não constituem o caminho ideal.

Surge assim a necessidade de encontrar procedimentos robustos, considerando que um procedimento é robusto se possuir boas propriedades quando o modelo suposto é válido e se, para além disso, não for muito sensível a pequenos afastamentos das hipóteses do modelo. Desta forma, um procedimento robusto deve ter boas propriedades, quer no modelo que se assume verdadeiro, quer nas suas vizinhanças.

3.2 Conceitos de robustez

Para concretizar a noção de robustez existem três abordagens fundamentais (*vide* Huber (1981)): a abordagem qualitativa, que se baseia na noção de continuidade, a abordagem quantitativa, que se baseia no conceito de ponto de ruptura (ou *breakdown point*) e a abordagem infinitesimal, baseada no conceito de função de influência. A cada uma das abordagens corresponde um critério de avaliação da robustez.

Antes de formalizar as definições relativas a cada uma das abordagens, introduz-se a noção de funcional estatístico, que facilita muito a sistematização matemática da

robustez. Para simplificar a exposição, considere-se uma população X, da qual se retira uma amostra aleatória $(X_1, ..., X_n)$. Seja $\mathfrak{X}$ o espaço amostral e $\mathfrak{F}(\mathfrak{X})$ uma família de *f.d.p.* definidas sobre $\mathfrak{X}$.

Definição 3.2.1. Um *funcional* estatístico T é uma aplicação que a cada elemento de $\mathfrak{F}(\mathfrak{X})$ faz corresponder um número real, ou um vector de números reais.

$\Diamond$

Quando as distribuições dependem de parâmetros, a representação desses parâmetros pode ser considerada através de funcionais estatísticos.

Como exemplo, considere-se que a variável aleatória X segue uma *f.d.p.* F, a qual depende do parâmetro desconhecido $\mathrm{E}[g(X)] \in \mathbb{R}$, onde g representa uma função qualquer (mensurável). Este parâmetro pode ser representado através do funcional

$$\begin{array}{rccl} T\colon & D_T \subset \mathfrak{F}(\mathfrak{X}) & \longrightarrow & \mathbb{R} \\ & F & \longmapsto & T(F) = \int g(x)dF(x). \end{array}$$

Note-se que o domínio do funcional T é constituído por todos os elementos de $\mathfrak{F}(\mathfrak{X})$ para os quais T está definido e abrange *f.d.p.* com diferentes características, correspondentes a variáveis contínuas, discretas ou mistas. Deste modo, se F for uma *f.d.p.* de uma variável aleatória contínua, então

$$T(F) = \int g(x)dF(x) = \int g(x)f(x)dx,$$

onde f representa a função densidade de probabilidade de X. Caso F seja a *f.d.p.* de uma variável aleatória discreta com suporte S_X, então a representação funcional de T é da forma

$$T(F) = \int g(x)dF(x) = \sum_{x \in S_X} g(x)f(x).$$

Assim, é possível representar o parâmetro $\mathrm{E}[g(X)]$ através do funcional $T(F)$, para qualquer $F \in D_T$.

Em particular, se F_n denotar a função de distribuição empírica da amostra aleatória $(X_1, ..., X_n)$, o parâmetro $T(F) = \mathrm{E}[g(X)]$ pode ser estimado calculando o valor do funcional T na distribuição $F_n \in \mathfrak{F}(\mathfrak{X})$, ou seja,

$$T(F_n) = \int g(x)dF_n(x) = \frac{1}{n}\sum_{i=1}^{n} g(x_i).$$

O exemplo mais simples é o do funcional média, que corresponde a tomar g como a função identidade. Calculando o funcional na distribuição da população X, tem-se que

$$T(F) = \int x dF(x) = \mu;$$

efectuando o cálculo do mesmo funcional na distribuição empírica, obtém-se a média amostral, pois a função de distribuição empírica atribui o peso $1/n$ a cada observação x_i. Logo,

$$T(F_n) = \int x dF_n(x) = \sum_{i=1}^{n} x_i \times \frac{1}{n} = \overline{x}.$$

A representação funcional de estimadores apresenta algumas limitações, dado que nem todos os estimadores podem ser representados por um funcional. No entanto, a grande maioria dos estimadores que são frequentemente utilizados em estatística têm uma representação funcional. Noutros casos, existe uma representação funcional que é assintoticamente equivalente ao estimador $T_n(X_1, ..., X_n)$, isto é, existe um funcional $T(F_n)$ tal que $T_n(X_1, ..., X_n) \neq T(F_n)$, mas a diferença $T_n(X_1, ..., X_n) - T(F_n) \to 0$ quando n tende para infinito. Isso acontece com a variância amostral corrigida, que não pode ser expressa como um funcional da função de distribuição empírica, mas que é assintoticamente equivalente à variância amostral (não corrigida).

A grande vantagem das representações funcionais é o facto de evidenciarem que o valor do parâmetro depende da distribuição considerada. Recorde-se que a hipótese dos dados serem provenientes da distribuição F pode falhar e a distribuição verdadeira pode ser uma função $G \in \mathfrak{F}(\mathfrak{X})$, situada numa vizinhança de F. Este facto poderia alterar significativamente o valor do parâmetro.

Considerando a representação funcional dos estimadores, de seguida apresentam-se as definições de robustez, de acordo com cada um dos principais conceitos.

A ideia fundamental da robustez qualitativa reside no facto de considerar que, se a distribuição G se encontra próxima da F, então as estimativas $T(G_n)$ e $T(F_n)$ também devem estar próximas.

Definição 3.2.2. Seja F uma *f.d.p.* e d(.) uma distância adequada entre funções de distribuição. O funcional T diz-se *contínuo em* F se, para qualquer sucessão de funções de distribuição $\{G_n\}_{n\in\mathbb{N}}$ tais que $\mathrm{d}(G_n, F) \xrightarrow{n} 0$, se tem que $|T(G_n) - T(F)| \xrightarrow{n} 0$.

$\Diamond$

Definição 3.2.3. Se o funcional T for contínuo em F, então o estimador $T(F_n)$ diz-se *robusto no sentido qualitativo.*

$\Diamond$

Se T for um funcional contínuo, então $T(F_n)$ é consistente. Isto acontece porque, ao tomar $G_n = F_n$ e, como a função de distribuição empírica converge quase certamente para F, então a continuidade implica que $|T(F_n) - T(F)| \xrightarrow{n} 0$, *i.e.*, $T(F_n)$ é consistente.

Para avaliar a noção de proximidade entre funções de distribuição é necessária uma distância adequada. Uma das distâncias mais utilizadas é a distância de Kolmogorov, definida por

$$\mathrm{d}_K(F, G) = \sup_{x\in\mathbb{R}} |F(x) - G(x)|. \tag{3.2.1}$$

Apesar do conceito de estimador qualitativamente robusto ser importante de um ponto de vista teórico, ele não é muito utilizado na prática. O problema é que o tratamento matemático da continuidade de funcionais é muito delicado, pelo que tem sido pouco desenvolvido. Por isso, não será seguida essa abordagem durante o presente trabalho.

Considere-se agora a abordagem quantitativa da robustez. Nesta abordagem, a noção fundamental é a de ponto de ruptura, que avalia o desempenho dos estimadores nas vizinhanças da distribuição $F \in \mathfrak{F}(\mathfrak{X})$ considerada. Assim, chama-se vizinhança de contaminação de F de raio δ ao conjunto de distribuições

$$\mathcal{V}_\delta(F) = \{G : G = (1-\delta)F + \delta H, H \in \mathfrak{F}(\mathfrak{X})\},$$

onde $\delta \in [0, 1]$. As distribuições de $\mathcal{V}_\delta(F)$ representam todas as possibilidades de contaminação do modelo F por uma qualquer distribuição H, com probabilidade δ. Se δ for igual a 0, $\mathcal{V}_0(F)$ apenas contém a distribuição F e não existe a possibilidade

de contaminação; se δ for igual a 1, $\mathcal{V}_1(F)$ coincide com $\mathfrak{F}(\mathfrak{X})$, a qual inclui as piores contaminações possíveis que o estimador pode sofrer.

É necessário saber qual é o pior efeito que um elemento de $\mathcal{V}_\delta(F)$ pode causar no estimador de funcional T. Deste modo, define-se

$$b_{T,F}(\delta) = \sup_{G \in \mathcal{V}_\delta(F)} |T(G) - T(F)|,$$

como sendo o *enviesamento assintótico máximo* de T na vizinhança $\mathcal{V}_\delta(F)$. É de notar que $b_{T,F}(1)$ é o pior valor que o enviesamento assintótico pode ter. O ponto de ruptura é o maior valor de δ que faz com que $b_{T,F}(\delta)$ não seja o pior possível.

Definição 3.2.4. Nas condições referidas anteriormente, chama-se *ponto de ruptura* do funcional T ao valor

$$\varepsilon^*(T, F) = \sup\{\delta \in [0, 1] : b_{T,F}(\delta) < b_{T,F}(1)\}.$$

$\Diamond$

O ponto de ruptura traduz a proporção máxima de contaminação por observações atípicas que um determinado estimador consegue suportar, sem devolver resultados absurdos. Nesse sentido, interessa que um estimador tenha um ponto de ruptura não nulo; e o estimador é tanto mais robusto quanto maior for o seu ponto de ruptura. Por outro lado, o maior valor que o ponto de ruptura pode tomar é 1/2, uma vez que não faz sentido que seja assumido um modelo, quando mais de metade das observações estão em desacordo com esse modelo.

A definição de ponto de ruptura não é simples e, em termos práticos, utiliza-se uma sua versão amostral.

Definição 3.2.5. O *ponto de ruptura empírico* de um estimador T_n de um parâmetro $\theta \in \mathbb{R}$, calculado numa realização amostral $(x_1, ..., x_n)$ é definido por

$$\varepsilon_n^*(T_n) = \frac{1}{n} \max \left\{ m : \max_{i_1,...,i_m} \sup_{y_1,...,y_m} |T_n(z_1, ..., z_n)| < \infty \right\},$$

onde $(z_1, ..., z_n)$ corresponde a uma amostra em que m observações $x_{i_1}, ..., x_{i_m}$ originais são substituídas por m valores arbitrários $y_1, ..., y_m$.

$\Diamond$

Esta definição foi apresentada para o caso mais simples de um parâmetro $\theta \in \Theta = \mathbb{R}$. No entanto, ela pode ser facilmente generalizada para outros espaços de parâmetro Θ.

Apesar do ponto de ruptura empírico ser determinado através de uma realização amostral, o seu valor raramente depende dessa realização. Em geral, ε_n^* depende apenas do número de observações consideradas e não depende da distribuição F. Quando n tende para infinito, geralmente ε_n^* tende para $\varepsilon^*(T, F)$. O ponto de ruptura empírico tem vindo a ser cada vez mais utilizado, uma vez que envolve conceitos simples e que não depende da distribuição de probabilidades subjacente.

Por último, resta apresentar a noção de robustez infinitesimal. A avaliação da robustez infinitesimal é feita através de um dos conceitos mais importantes para a robustez, designadamente o de função de influência.

Definição 3.2.6. Seja F uma *f.d.p.* pertencente ao domínio do funcional T e seja $F_{x,\delta}$ a *f.d.p.* contaminada em x que tem a forma

$$F_{x,\delta} = (1 - \delta)F + \delta\Delta_x, \tag{3.2.2}$$

onde Δ_x representa a função de distribuição com massa unitária em x. Chama-se *função de influência* de T em F a

$$IF(x; T, F) = \lim_{\delta \to 0} \frac{T(F_{x,\delta}) - T(F)}{\delta},$$

em todo o x onde o limite existe.

$\Diamond$

Note-se que a quantidade $[T(F_{x,\delta}) - T(F)]/\delta$ representa a taxa de variação média do funcional T quando a distribuição F é sujeita a uma proporção δ de contaminação. Por isso, a função de influência pode ser interpretada como sendo uma derivada do funcional que define o estimador. Para que o estimador seja robusto no sentido infinitesimal, é necessário que o funcional que o define não varie muito quando a distribuição está contaminada. Isso significa que a sua função de influência é limitada.

Definição 3.2.7. Seja F uma *f.d.p.* e T um funcional com função de influência $IF(x; T, F)$. Chama-se *sensibilidade a grandes erros* a

$$\gamma^*(T; F) = \sup_x |IF(x; T, F)|.$$

Diz-se que T é *B-robusto* em F se $\gamma^*(T;F) < \infty$.

$\Diamond$

A sensibilidade a grandes erros representa a perturbação máxima que uma contaminação infinitesimal δ pode causar no estimador definido pelo funcional T. Quanto maior for $\gamma^*(T;F)$, mais sensível é o estimador a erros grosseiros. Se a sensibilidade a grandes erros for infinita, então o estimador de funcional T não é robusto, de acordo com esta abordagem.

Seguidamente, mostra-se o processo de cálculo da função de influência para a média e para a mediana.

Exemplo 3.2.1. Pretende-se determinar a função de influência do estimador média amostral, o qual é definido por

$$T_n = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

e que, como já foi referido, pode ser representado à custa do funcional $\mu(F) = \int u dF(u)$.

O valor do funcional na distribuição contaminada (3.2.2) é dado por

$$\begin{aligned}
\mu(F_{x,\delta}) &= \mu((1-\delta)F + \delta\Delta_x) \\
&= \int u d((1-\delta)F(u) + \delta\Delta_x(u)) \\
&= (1-\delta)\int u dF(u) + \delta \int u d\Delta_x(u) \\
&= (1-\delta)\mu(F) + \delta x \\
&= \mu(F) + \delta(x - \mu(F)).
\end{aligned}$$

Consequentemente, vem que

$$IF(x;\mu,F) = \lim_{\delta\to 0} \frac{\mu(F_{x,\delta}) - \mu(F)}{\delta} = \lim_{\delta\to 0} \frac{\mu(F) + \delta(x-\mu(F)) - \mu(F)}{\delta} = x - \mu(F).$$

A Figura 3.1 apresenta o gráfico da função de influência da média amostral.

Concluindo, se o suporte da variável aleatória X (que tem *f.d.p.* F) for ilimitado, então a função de influência da média amostral é ilimitada. Então, a média amostral não é um estimador robusto do ponto de vista infinitesimal.

$\Diamond$

Exemplo 3.2.2. Considere-se agora que se pretende determinar a função de influência do estimador mediana, para uma população com *f.d.p.* F e função densidade f. Para

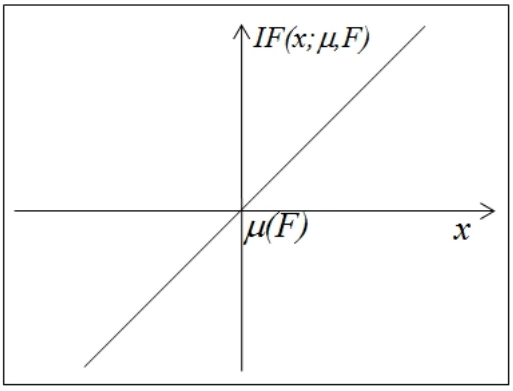

Figura 3.1: Representação do gráfico da função de influência da média amostral.

simplificar a apresentação, supõe-se que F é contínua e invertível no intervalo $]0,1[$. Sendo assim, não existe nenhuma ambiguidade quanto à definição da função inversa F^{-1}, e a mediana é representada através do funcional

$$m(F) = F^{-1}\left(\frac{1}{2}\right).$$

Como a distribuição contaminada $F_{x,\delta}$ pode ser escrita na forma

$$F_{x,\delta}(u) = \begin{cases} (1-\delta)F(u) & \text{se} \quad u < x \\ (1-\delta)F(u) + \delta & \text{se} \quad u \geq x \end{cases},$$

calculando $F_{x,\delta}$ no ponto $m(F)$, é possível concluir que, se $m(F) < x$, então $m(F_{x,\delta}) > m(F)$; e que se $m(F) > x$, então $m(F_{x,\delta}) < m(F)$. Assim, a mediana de $F_{x,\delta}$ depende da relação entre o ponto x e a mediana de F. De facto, se $x > m(F)$ então

$$F_{x,\delta}(m(F_{x,\delta})) = \frac{1}{2} \Leftrightarrow (1-\delta)F(m(F_{x,\delta})) = \frac{1}{2} \Leftrightarrow m(F_{x,\delta}) = F^{-1}\left(\frac{1/2}{1-\delta}\right) (> m(F));$$

caso $x < m(F)$ então

$$F_{x,\delta}(m(F_{x,\delta})) = \frac{1}{2} \Leftrightarrow (1-\delta)F(m(F_{x,\delta}))+\delta = \frac{1}{2} \Leftrightarrow m(F_{x,\delta}) = F^{-1}\left(\frac{1/2-\delta}{1-\delta}\right) (< m(F)).$$

Como a função de influência pode ser determinada por

$$IF(x; m, F) = \frac{\partial}{\partial\delta} m(F_{x,\delta})\Big|_{\delta=0},$$

e como

$$\frac{\partial}{\partial\delta} m(F_{x,\delta}) = \begin{cases} \frac{-1/2-2\delta}{(1-\delta)^2}\left[f\left(F^{-1}\left(\frac{1/2-\delta}{1-\delta}\right)\right)\right]^{-1} & \text{se} \quad x < m(F) \\ \frac{1/2}{(1-\delta)^2}\left[f\left(F^{-1}\left(\frac{1/2}{1-\delta}\right)\right)\right]^{-1} & \text{se} \quad x > m(F) \end{cases},$$

então, tomando $\delta = 0$, obtém-se

$$IF(x; m, F) = \begin{cases} -\frac{1}{2f(m(F))} & \text{se} \quad x < m(F) \\ \frac{1}{2f(m(F))} & \text{se} \quad x > m(F) \end{cases} .$$

A Figura 3.2 apresenta o gráfico da função de influência da mediana amostral.

Figura 3.2: Representação do gráfico da função de influência da mediana amostral.

Assim, a mediana é um estimador que tem função de influência limitada.

◊

3.3 Estimação robusta

A dificuldade em conciliar a robustez com a eficiência em modelos normais é bem conhecida. De facto, geralmente existe um conflito entre a robustez e a eficiência – quando se aumenta a robustez, diminui-se a eficiência, e quando se aumenta a eficiência, diminui-se a robustez. No seguimento deste capítulo, faz-se uma apresentação resumida de alguns estimadores que conseguem conciliar bem a robustez com a eficiência, considerando distribuições normais. Na apresentação, focam-se especialmente os estimadores que serão utilizados em capítulos seguintes. Porém, a presente secção, começa por relembrar algumas noções fundamentais em inferência estatística, de modo a que se possam entender os princípios que presidiram ao desenvolvimento dos estimadores robustos.

3.3.1 Questões de invariância

De entre as propriedades desejáveis num estimador, é importante que ele seja invariante em relação a certas transformações de dados, de modo a assegurar que as estimativas do parâmetro são modificadas de forma coerente com a transformação efectuada sobre os dados originais. Num contexto formal, a invariância é expressa através de diversas definições de equivariância, que aqui se apresentam depois de precisar as noções de modelo de localização e de modelo de escala.

Definição 3.3.1. Um *modelo de localização*, com *parâmetro de localização* μ, é uma família de distribuições $\{F_\mu, \mu \in \mathbb{R}\}$ tal que

$$F_\mu(x) = F(x - \mu),$$

onde $F = F_0$ representa uma *f.d.p.* univariada que define a família do modelo.

$\Diamond$

Definição 3.3.2. Um *modelo de escala*, com *parâmetro de escala* σ, é uma família de distribuições $\{F_\sigma, \sigma \in \mathbb{R}^+\}$ tal que

$$F_\sigma(x) = F\left(\frac{x}{\sigma}\right),$$

onde $F = F_1$ representa uma *f.d.p.* univariada que define a família do modelo.

$\Diamond$

Por exemplo, no modelo de localização com distribuição normal, o parâmetro de localização é a média, que é uma medida de localização. No modelo de escala com a mesma distribuição, o parâmetro de escala é o desvio padrão, que é uma medida de escala. Se se fizer uma translação ao conjunto de dados, então espera-se que a medida de localização também sofra essa mesma translação. Por outro lado, quando se muda a escala do conjunto de dados, é de esperar que a medida de dispersão também sofra o mesmo efeito.

Associado aos modelos de localização e de escala encontram-se os estimadores dos respectivos parâmetros. Os estimadores do modelo de localização podem ter as propriedades de invariância que a seguir se definem.

Definição 3.3.3. Um estimador T_n de um parâmetro de localização diz-se *equivariante em relação à localização* se para qualquer amostra aleatória $(X_1, ..., X_n)$,

$$T_n(X_1 + k, ..., X_n + k) = T_n(X_1, ..., X_n) + k, \quad \forall_{k \in \mathbb{R}},$$

e diz-se *equivariante em relação à escala* se

$$T_n(kX_1, ..., kX_n) = kT_n(X_1, ..., X_n), \quad \forall_{k \in \mathbb{R} \setminus \{0\}}.$$

$\Diamond$

As propriedades análogas para o modelo de escala são as seguintes:

Definição 3.3.4. Um estimador S_n de um parâmetro de escala diz-se *equivariante em relação à localização* se para qualquer amostra aleatória $(X_1, ..., X_n)$,

$$S_n(X_1 + k, ..., X_n + k) = S_n(X_1, ..., X_n), \quad \forall_{k \in \mathbb{R}},$$

e diz-se *equivariante em relação à escala* se

$$S_n(kX_1, ..., kX_n) = kS_n(X_1, ..., X_n), \quad \forall_{k \in \mathbb{R}^+}.$$

$\Diamond$

É desejável que os estimadores de localização e de escala verifiquem as respectivas propriedades de equivariância.

Também existem modelos de localização e escala em que, simultaneamente, são consideradas as duas características anteriores.

Definição 3.3.5. Um *modelo de localização e escala*, com *parâmetro de localização* μ e *parâmetro de escala* σ, é uma família de distribuições $\{F_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$, com $\boldsymbol{\theta} = (\mu, \sigma) \in \Theta = \mathbb{R} \times \mathbb{R}^+$, tal que

$$F_{\boldsymbol{\theta}}(x) = F\left(\frac{x - \mu}{\sigma}\right),$$

onde $F = F_{0,1}$ representa uma *f.d.p.* univariada que define a família do modelo.

$\Diamond$

A noção de invariância também se adapta a modelos de regressão.

Definição 3.3.6. Considere-se o modelo de regressão linear

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon, \tag{3.3.1}$$

onde $\mathbf{X}$ é uma matriz $(n \times p)$ de observações dos regressores, $\mathbf{y} \in \mathbb{R}^n$ é um vector de observações da variável resposta, $\boldsymbol{\beta} \in \mathbb{R}^p$ é um vector de parâmetros e $\epsilon \in \mathbb{R}^n$ é o vector dos erros do modelo. Seja $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ um estimador do parâmetro $\boldsymbol{\beta}$. Então

$\hat{\boldsymbol{\beta}}$ diz-se um *estimador equivariante em relação à regressão* se

$$\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y} + \mathbf{X}\lambda) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}) + \lambda, \quad \forall_{\lambda \in \mathbb{R}^p};$$

$\hat{\boldsymbol{\beta}}$ diz-se um *estimador equivariante em relação à escala* se

$$\hat{\boldsymbol{\beta}}(\mathbf{X}, k\mathbf{y}) = k\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}), \quad \forall_{k \in \mathbb{R}};$$

$\hat{\boldsymbol{\beta}}$ diz-se um *estimador equivariante afim* se

$$\hat{\boldsymbol{\beta}}(\mathbf{AX}, \mathbf{y}) = \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}),$$

para qualquer matriz $\mathbf{A}$ não singular de dimensão $(p \times p)$.

$\Diamond$

Note-se que, quando se transforma a variável resposta $\mathbf{y}$ em $\mathbf{y}^*$ por uma transformação da forma

$$\mathbf{y}^* = \mathbf{y} + \mathbf{X}\lambda,$$

para algum $\lambda \in \mathbb{R}^p$, então o novo modelo verifica a equação $\mathbf{y}^* = \mathbf{X}(\boldsymbol{\beta} + \lambda) + \epsilon$. Por isso, o parâmetro $\boldsymbol{\beta}^*$ do modelo transformado passa a ser $\boldsymbol{\beta}^* = \boldsymbol{\beta} + \lambda$. Assim, é desejável que as estimativas se comportem coerentemente, isto é, que quando se transformam os dados de acordo com a transformação anterior, o estimador verifique $\hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}^*) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y} + \mathbf{X}\lambda) = \hat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{y}) + \lambda$. É possível utilizar um raciocínio análogo para explicar o interesse das outras duas propriedades de equivariância do modelo de regressão linear.

3.3.2 Questões de robustez e de eficiência

Supondo conhecidas as distribuições de probabilidade associadas aos modelos estatísticos, os parâmetros da distribuição são frequentemente estimados através do método

de máxima verosimilhança, devido às propriedades dos resultantes estimadores. No caso da distribuição normal, é bem conhecido que os estimadores de máxima verosimilhança da média e do desvio padrão são, respectivamente, a média amostral e o desvio padrão amostral. Estes estimadores são equivariantes, são consistentes e eficientes sob o modelo normal; no entanto, não apresentam boas propriedades no que diz respeito à robustez. De facto, ambos os estimadores têm função de influência ilimitada e ponto de ruptura nulo. Por isso, não são robustos, e as estimativas que produzem não conseguem resistir a desvios das hipóteses do modelo, sendo particularmente sensíveis a observações atípicas.

Por outro lado, existem estimadores que são muito robustos mas que são pobres no que respeita à eficiência, particularmente em modelos normais. Como exemplo, considere-se a mediana amostral no modelo de localização normal. Como se mostrou no *Exemplo* 3.2.2, a mediana tem função de influência limitada e, por isso, é um estimador B-robusto; tem sensibilidade a grandes erros $\gamma^*(m;\Phi) = (\pi/2)^{1/2} \approx 1.253$, que é o menor valor em estimadores do parâmetro de localização no modelo normal (*vide* Hampel, Ronchetti, Rousseeuw e Stahel (1986)); por fim ainda tem um ponto de ruptura igual a 1/2, o que significa que a mediana amostral devolve estimativas razoáveis sempre que existam menos de metade de observações da amostra que sejam atípicas. Contudo, a mediana amostral tem uma eficiência assintótica em modelos normais de 63.7%, que é um valor baixo. Por isso, o que se perde em termos de eficiência, não é compensado com o que se pode vir a ganhar em termos de robustez.

Referindo também o caso do modelo de escala, considere-se o desvio absoluto mediano, denotado abreviadamente por *MAD*. Tal como a mediana no modelo de localização, o *MAD* tem uma função de influência limitada e um ponto de ruptura igual a 1/2. A sua sensibilidade a grandes erros é de $\gamma^*(MAD;\Phi) = 1.166$, que também é o menor valor que se pode obter em estimadores do parâmetro de escala no modelo normal (*vide* Hampel *et al.* (1986)). Porém, o *MAD* tem uma eficiência assintótica em modelos normais de apenas 37%. Mais uma vez, o "custo"da robustez é demasiado elevado.

Surgiu assim a necessidade de procurar estimadores robustos, mantendo a eficiência em níveis aceitáveis, ao trabalhar em modelos normais. De seguida, apresentam-se

propostas da literatura, que conseguem conciliar boas propriedades de robustez com boa eficiência em modelos normais.

3.3.3 Os estimadores-MM

Os estimadores-MM foram propostos por Yohai (1987) e, actualmente, são considerados os estimadores de localização que melhor conseguem conciliar a robustez com a eficiência em modelos normais. O nome MM resulta do facto dos estimadores serem construídos com base em dois estimadores-M, um de localização e outro de escala. Por isso, é necessário introduzir o conceito geral de estimador-M para, de seguida, precisar a definição de estimador-MM.

Assim, considere-se que $(X_1, ..., X_n)$ é uma amostra aleatória da população X, cuja distribuição depende de um parâmetro de localização μ desconhecido. A ideia inicial foi definir o *estimador-M* pela solução da equação

$$T_n = \arg\min_{\mu\in\Theta} \sum_{i=1}^{n} \rho(X_i; \mu), \tag{3.3.2}$$

para uma função ρ conveniente, definida em $\mathfrak{X} \times \Theta$.

Repare-se que (3.3.2) corresponde a uma generalização dos estimadores de máxima verosimilhança, os quais se obtêm tomando $\rho = -\log f$, onde f é a função densidade de X. Foi devido a esse facto que Huber (1964) lhes deu o nome de estimadores-M. Consequentemente, supondo as condições de regularidade usuais, qualquer estimador de máxima verosimilhança é um estimador-M (mas a implicação recíproca não é verdadeira).

Neste contexto, a função ρ designa-se por função objectivo. Em geral, a função objectivo satisfaz as seguintes propriedades:

- é não negativa, ou seja, $\forall_{x\in\mathbb{R}}\ \rho(x) \geq 0$;
- $\rho(0) = 0$;
- é par, *i.e.*, $\rho(x) = \rho(-x),\ \forall_{x\in\mathbb{R}}$;
- é não decrescente para todo o $x > 0$;
- é uma função contínua.

No seguimento, admite-se que uma função objectivo goza das propriedades anteriores. Se existir a derivada de ρ, denotada por $\psi(x;\mu) = (\partial/\partial\mu)\rho(x;\mu)$, então o estimador-M também pode ser definido como o estimador T_n que satisfaz a equação

$$\sum_{i=1}^{n} \psi(X_i; T_n) = 0. \tag{3.3.3}$$

A função ψ que consta na equação anterior é conhecida por *função ψ do estimador-M* e é de grande importância na construção de estimadores robustos. Efectivamente, pode provar-se que a função de influência de um estimador-M é proporcional à sua função ψ. Logo, partindo de funções ψ limitadas, é possível desenvolver estimadores robustos por construção (veja-se, por exemplo, em Hampel *et al.* (1986)). Por esse motivo, e porque a função ψ facilita a generalização a estimadores multivariados, a equação (3.3.3) é geralmente preferida para definir os estimadores-M.

No caso particular do modelo de localização, as condições (3.3.2) e (3.3.3) podem ser, respectivamente, substituídas por

$$T_n = \arg\min_{\mu\in\Theta} \sum_{i=1}^{n} \rho(X_i - \mu)$$

e por

$$\sum_{i=1}^{n} \psi(X_i - T_n) = 0.$$

Através da equação anterior, é imediato verificar que os estimadores-M são equivariantes em relação à localização. No entanto, o mesmo já não acontece para a equivariância em relação à escala. De facto, esta última propriedade não se verifica automaticamente para qualquer função ρ.

Para tornar os estimadores-M de localização equivariantes em relação à escala, foi necessário considerar um estimador auxiliar de escala, S_n, e definir o estimador-M de localização pela solução de

$$T_n = \arg\min_{\mu\in\Theta} \sum_{i=1}^{n} \rho\left(\frac{X_i - \mu}{S_n}\right).$$

É fácil verificar que, se S_n for um estimador equivariante em relação à escala, então kT_n, com $k > 0$, é solução da equação que se obtém substituindo os $X_1, ..., X_n$ por $kX_1, ..., kX_n$ na equação anterior.

De acordo com Hampel *et al.* (1986), o estimador de escala S_n deve ser o mais robusto possível, não sendo muito importante a sua eficiência. Note-se que o interesse está em estimar o parâmetro de localização μ e, neste caso, S_n é apenas um estimador auxiliar.

A escolha do estimador auxiliar S_n dá origem a diferentes classes de estimadores--M. Como se verá de seguida, a classe dos estimadores-MM é caracterizada por utilizar S_n como sendo aquele que produz a menor estimativa de entre todos os estimadores--M de escala. Assim, para apresentar a definição de estimador-MM de localização, é necessário começar por definir um estimador-M de escala.

Definição 3.3.7. Um estimador $\hat{\sigma}_n$ diz-se um *estimador-M de escala* se satisfaz uma equação da forma

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{X_i-\hat{\mu}}{\hat{\sigma}_n}\right)=b, \tag{3.3.4}$$

onde $\hat{\mu}$ é uma estimativa obtida através de um estimador equivariante em relação à localização e à escala, e b é uma constante de afinação positiva.

$\Diamond$

Repare-se que, para que a equação (3.3.4) tenha solução, é necessário que $0 < b < \lim_{x\to\infty}\rho(x)$. Costuma-se tomar $b = \mathrm{E}[\rho(X)]$. Geralmente assume-se que ρ é uma função limitada e que, sem perda de generalidade, $\lim_{x\to\infty}\rho(x) = 1$. Seguidamente assumem-se estas duas condições e, por isso, supõe-se que $b \in]0,1[$.

A menor estimativa obtida com estimadores-M de escala é um estimador designado por estimador-S de escala, o qual foi proposto por Rousseeuw e Yohai (1984).

Definição 3.3.8. Considerem-se, para qualquer $t \in \mathbb{R}$, os resíduos $X_1 - t, ..., X_n - t$, onde t é uma estimativa de localização e $\hat{\sigma}_n(t)$ é a respectiva estimativa-M de escala, que satisfaz a equação

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{X_i-t}{\hat{\sigma}_n(t)}\right)=b, \quad b\in]0,1[.$$

O *estimador-S de escala* é definido por

$$S_n=\inf_{t\in\mathbb{R}}\hat{\sigma}_n(t).$$

$\Diamond$

Finalmente, é possível apresentar a definição de estimador-MM.

Definição 3.3.9. T_n diz-se um *estimador-MM de localização* se

$$T_n = \arg\min_{\mu\in\Theta} \sum_{i=1}^{n} \rho_1\left(\frac{X_i-\mu}{S_n}\right),$$

onde S_n é um estimador-S de escala, isto é, onde S_n minimiza o estimador-M de escala $\hat{\sigma}_n(t)$ implicitamente definido pela equação

$$\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{X_i-t}{\hat{\sigma}_n(t)}\right) = b, \quad b \in]0,1[.$$

$\Diamond$

Note-se que, quer o ponto de ruptura, quer a eficiência dos estimadores-MM, são determinados pelas funções objectivo (ρ_0 e ρ_1) e pela constante de afinação b. Yohai (1987) mostrou que, para que os estimadores-MM sejam consistentes, é essencial que $\rho_1(x)$ seja menor ou igual a $\rho_0(x)$, para qualquer $x \in \mathbb{R}$.

Em geral, as funções ρ_0 e ρ_1 pertencem à família *bi-square* proposta por Beaton e Tukey (1974), a qual é da forma

$$\rho(x) = \min\left\{1, 1-\left(1-x^2\right)^3\right\}. \tag{3.3.5}$$

Assim, geralmente toma-se $\rho_0(x) = \rho(x/c_0)$ e $\rho_1(x) = \rho(x/c_1)$. Para que a função ρ_1 nunca seja superior a ρ_0, toma-se $c_1 \geq c_0$. Quando a população X segue uma distribuição normal, utiliza-se $c_0 = 1.56$ e $b = 0.5$, para garantir a consistência do estimador-S de escala. O valor da constante c_1 permite regular a eficiência do estimador – quanto maior for a constante c_1, maior será a eficiência assintótica do estimador-MM. A Tabela 3.1 foi retirada de Maronna, Martin e Yohai (2006) e representa a relação entre os valores de c_1 e a eficiência assintótica do estimador-MM, supondo o modelo de localização normal.

Os estimadores-MM podem ser construídos de modo a atingir até 95% de eficiência assintótica, quando X segue uma distribuição normal. Por outro lado, como se pode ver em Maronna *et al.* (2006), nas condições anteriores, o ponto de ruptura do estimador-MM é igual ao mínimo entre b e $1-b$. Por isso, quando se utiliza a constante de afinação $b = 0.5$, garante-se que o estimador-MM tem ponto de ruptura máximo.

Eficiência	80%	85%	90%	95%
c_1	3.14	3.44	3.88	4.68

Tabela 3.1: Relação entre a constante de afinação c_1 da função objectivo e a eficiência assintótica do estimador-MM, no modelo de localização normal.

Consequentemente, os estimadores-MM conseguem devolver estimativas credíveis sempre que existam menos de metade de observações da amostra que sejam atípicas. Por outro lado, Yohai (1987) também mostrou que estes estimadores são consistentes e que têm uma distribuição assintótica normal. Estas propriedades fazem com que os estimadores-MM sejam dos estimadores de localização que melhor conseguem conciliar a robustez com a eficiência em modelos normais.

Quando se fala dos estimadores-MM, não se pode deixar de os considerar no contexto da regressão, uma vez que foi aí que Yohai (1987) os apresentou. A definição é análoga à dos estimadores-MM de localização.

Definição 3.3.10. Considere-se o modelo de regressão linear referido em (3.3.1) onde, como é usual, $\mathbf{y} = [y_1 \ ... \ y_n]^{\mathrm{T}}$ e $\mathbf{X} = [\mathbf{x}_1 \ ... \ \mathbf{x}_n]^{\mathrm{T}}$.

O *estimador-MM* de $\boldsymbol{\beta} \in \mathbb{R}^p$, que se denota por $\hat{\boldsymbol{\beta}}_{MM}$, é definido como solução da equação vectorial

$$\sum_{i=1}^{n} \rho_1' \left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{MM}}{\hat{\sigma}_n} \right) \mathbf{x}_i = \mathbf{0},$$

onde, na j-ésima equação, ρ_1' representa a derivada da função ρ_1 relativamente à componente $\beta_j \in \boldsymbol{\beta}$, para $j = 1, ..., p$, e $\hat{\sigma}_n$ é a estimativa do estimador-S de escala, a qual representa o valor mínimo de $\hat{\sigma}_n(\boldsymbol{\beta})$ que satisfaz a equação

$$\frac{1}{n} \sum_{i=1}^{n} \rho_0 \left(\frac{y_i - \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}}{\hat{\sigma}_n(\boldsymbol{\beta})} \right) = b,$$

onde b é a constante de afinação.

$\Diamond$

Yohai (1987) mostrou que $\hat{\boldsymbol{\beta}}_{MM}$ goza das mesmas propriedades já referidas para o estimador-MM de localização. Assim, considerando as mesmas funções objectivo e a mesma constante de afinação que no modelo de localização, os estimadores-MM de

regressão também conseguem possuir, simultaneamente, um ponto de ruptura igual a 1/2 e uma eficiência assintótica de 95%, supondo a distribuição normal dos erros do modelo. Por outro lado, como se pode ver em Maronna *et al.* (2006), os estimadores-MM também gozam das propriedades de equivariância apresentadas na **Definição 3.3.6**.

Repare-se que a estimativa $\hat{\boldsymbol{\beta}}_{MM}$ pode não corresponder ao ponto de mínimo absoluto da função ρ_1. Contudo, Yohai (1987) mostrou que, quando a função ρ_1 nunca é superior a ρ_0, e quando os erros do modelo são normais, então existem mínimos locais que devolvem estimativas com as mesmas boas propriedades que a estimativa do mínimo absoluto de ρ_1. Para que isso aconteça, basta que o valor do mínimo local seja inferior ao valor de ρ_1 no ponto $\boldsymbol{\beta}$ que minimiza $\hat{\sigma}_n(\boldsymbol{\beta})$. Por isso, não é necessário encontrar exactamente o mínimo absoluto de ρ_1. Esta propriedade é importante uma vez que facilita bastante o cálculo dos estimadores-MM.

3.3.4 O estimador Q_n

Tal como acontece com os estimadores-MM, o estimador Q_n proposto por Rousseeuw e Croux (1993) é um dos estimadores que melhor consegue conciliar a robustez com a eficiência mas, neste caso, em modelos de escala normais.

Considere-se uma amostra aleatória $(X_1, ..., X_n)$ de uma população X de parâmetro de escala σ desconhecido.

Definição 3.3.11. O *estimador* Q_n do parâmetro de escala σ é definido por

$$Q_n(X_1, ..., X_n) = c \times \{|X_i - X_j| : 1 \leq i < j \leq n\}_{(k)}, \tag{3.3.6}$$

onde $k = \binom{[n/2]+1}{2}$, o índice (k) representa a estatística de ordem k do conjunto considerado e c é uma constante que torna o estimador Q_n consistente.

$\Diamond$

Rousseeuw e Croux (1993) mostraram que, se X seguir uma distribuição normal, o estimador Q_n é consistente quando a constante c toma o valor $\frac{1}{\sqrt{2}\Phi^{-1}(5/8)} \approx 2.2219$, onde Φ designa a $f.d.p.$ normal standardizada.

O estimador Q_n é apenas assintoticamente centrado. Por isso, se n for pequeno, deve-se utilizar um factor de correcção do enviesamento. Croux e Rousseeuw (1992)

apresentaram os factores de correcção do enviesamento do estimador Q_n para valores de n inferiores a 40.

Em Rousseeuw e Croux (1993), os autores mostram que, em modelos normais, este estimador tem uma eficiência assintótica de 82%, um ponto de ruptura de 1/2 e uma função de influência limitada, com sensibilidade a grandes erros de $\gamma^*(Q_n; \Phi) = 2.069$. Para além disso, os mesmos autores confirmaram que, nas condições assumidas nesta subsecção, o estimador Q_n tem uma distribuição assintótica normal. Portanto, o estimador Q_n é uma boa opção para conciliar a robustez com a eficiência em modelos normais de escala.

Capítulo 4

Adaptação da metodologia *bootstrap* a processos geoestatísticos

Neste capítulo apresenta-se a metodologia *bootstrap* tradicional e as suas extensões a estruturas de dependência temporais. De seguida, propõe-se uma metodologia de reamostragem para estruturas espaciais. Essa metodologia foi inspirada pelo *bootstrap* por blocos circulares, o qual já é utilizado em séries temporais, uma vez que esta versão de *bootstrap* preserva características de dependência. Faz-se um estudo sobre o enviesamento da média amostral de um processo geoestatístico com estacionaridade forte, para o esquema de reamostragem proposto.

4.1 O princípio da metodologia *bootstrap*

A metodologia *bootstrap* clássica foi proposta por Efron (1979) num contexto de observações *i.i.d.*. Ela tem por base um processo de reamostragem computacional, automático e intensivo, que permite a resolução de um grande número de problemas de inferência estatística.

Considere-se que se pretende estudar a distribuição de uma certa estatística, em particular, de um estimador $\hat{\theta} = T(\mathbf{X})$, onde $\mathbf{X} = (X_1, ..., X_n)$, a partir de uma realização de uma amostra aleatória, $\mathbf{x} = (x_1, ..., x_n)$. A distribuição de $\hat{\theta}$ depende da distribuição da população, mas em diversas situações não se conhece a *f.d.p.* F da população; outras vezes, embora se conheça a *f.d.p.* F, a distribuição da estatística é muito difícil, ou mesmo impossível, de tratar matematicamente. Por isso, Efron (1979)

propôs que se estudasse a distribuição de $\hat{\theta}$, aproximando F por uma sua estimativa $\hat{F}$, a qual é obtida através da realização amostral $\mathbf{x}$. A partir da população com *f.d.p.* $\hat{F}$, são geradas amostras aleatórias, encarando $\hat{F}$ como se fosse a verdadeira *f.d.p.* da população. Consequentemente, espera-se que a distribuição empírica da estatística $\hat{\theta}$ para esse conjunto de amostras, aproxime a verdadeira distribuição de $\hat{\theta}$ (de acordo com a distribuição F da população).

Resumidamente, a metodologia *bootstrap* consiste em:

- considerar que a estimativa $\hat{F}$ é a *f.d.p.* da população;
- formar novas amostras aleatórias, retirando observações de acordo com $\hat{F}$; estas novas amostras são designadas por amostras *bootstrap* e representam-se por $\mathbf{X}^{(b)*} = (X_1^{(b)*}, ..., X_n^{(b)*})$, sendo $\mathbf{x}^{(b)*} = (x_1^{(b)*}, ..., x_n^{(b)*})$ as suas realizações;
- para cada uma das amostras *bootstrap* geradas, calcular o valor da estatística, obtendo assim B realizações $\hat{\theta}^{(b)*} = T(\mathbf{x}^{(b)*})$, para $b = 1, ..., B$, as quais são designadas por réplicas *bootstrap*.

As B réplicas *bootstrap* aproximam a distribuição da estatística $\hat{\theta}$. O paralelismo entre a metodologia usual e a metodologia *bootstrap* encontra-se esquematizado na Figura 4.1.

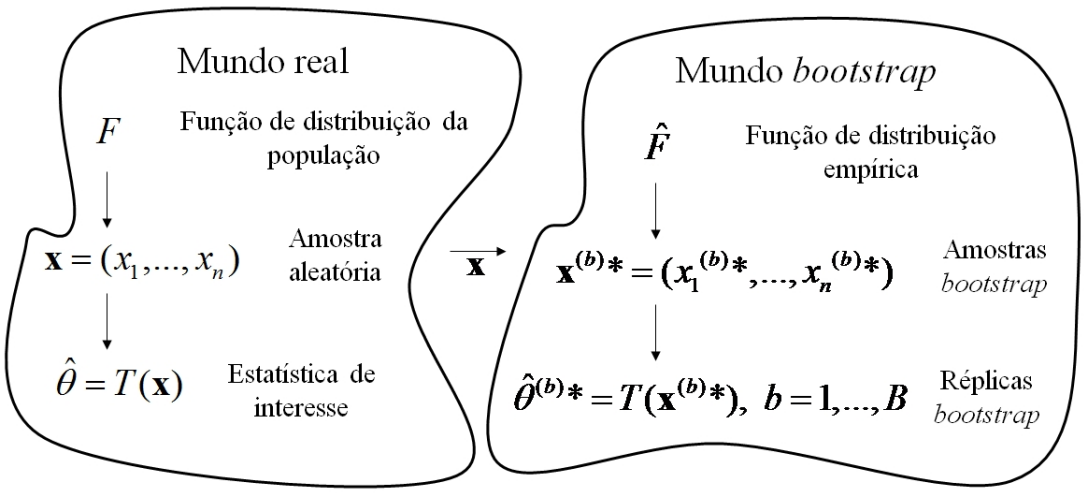

Figura 4.1: Esquema para ilustrar a reamostragem *bootstrap*.

Existem dois tipos fundamentais de *bootstrap*, o paramétrico e o não paramétrico. Ambos se distinguem pela função de distribuição $\hat{F}$ que utilizam para aproximar a *f.d.p.* F da população.

Para caracterizar o *bootstrap* paramétrico, represente-se por $F(x|\theta)$ a *f.d.p.* F, evidenciando a sua dependência em relação ao valor de θ. Quando θ é estimado pelo estimador de máxima verosimilhança, então o verdadeiro valor de θ é aproximado pela estimativa $\hat{\theta}_{MV}$. O *bootstrap* paramétrico assume que $\hat{F}(x) = F(x|\hat{\theta}_{MV})$. Então, as amostras *bootstrap* são geradas aleatoriamente a partir da distribuição $F(x|\hat{\theta}_{MV})$.

O *bootstrap* não paramétrico caracteriza-se por considerar que $\hat{F}$ é a própria função de distribuição empírica da amostra $\mathbf{x}$. Por isso as amostras *bootstrap* são retiradas aleatoriamente e com reposição, directamente da amostra original.

De um ponto de vista prático, a principal diferença entre os dois tipos de *bootstrap* reside no facto do *bootstrap* paramétrico exigir que se assuma uma *f.d.p.* F para a população e que se conheça o estimador de máxima verosimilhança de θ. Por isso, o método com mais adeptos é o *bootstrap* não paramétrico. Ao longo deste trabalho, seguir-se-á a metodologia *bootstrap* não paramétrica.

No *bootstrap* não paramétrico de Efron, as amostras *bootstrap* são obtidas por um processo de reamostragem simples, o qual é efectuado seleccionando, aleatoriamente e com reposição, as observações da amostra $\mathbf{x}$. Por outras palavras, cada elemento $x_i^{(b)*}$ de uma amostra *bootstrap* representa uma realização da variável aleatória X_j, componente da amostra original $(X_1, ..., X_n)$. Este esquema de reamostragem permite formar n^n amostras *bootstrap* com n componentes, obtidas a partir da amostra $\mathbf{x}$. Repare-se que algumas das amostras *bootstrap* só diferem na ordem por que estão dispostas as suas componentes.

Para exemplificar o estudo da distribuição de um estimador usando o *bootstrap* não paramétrico, considere-se o estudo do enviesamento de um estimador $\hat{\theta}$, supondo esse esquema de reamostragem.

Exemplo 4.1.1. Pretende-se estimar a esperança de um estimador $\hat{\theta}$ de um parâmetro θ usando reamostragem *bootstrap*. Para tal, começa por se considerar a amostra $\mathbf{x}$ como se fosse a população. Então, de $\mathbf{x}$ retiram-se aleatoriamente e com reposição as B amostras *bootstrap* $\mathbf{x}^{(b)*} = (x_1^{(b)*}, ..., x_n^{(b)*})$, para $b = 1, ..., B$.

Cada amostra *bootstrap* dá origem a uma réplica *bootstrap* de $\hat{\theta}$, ou seja,

$$\begin{array}{rcl} \mathbf{x}^{(1)*} = (x_1^{(1)*}, ..., x_n^{(1)*}) & \longrightarrow & \hat{\theta}^{(1)*}; \\ \mathbf{x}^{(2)*} = (x_1^{(2)*}, ..., x_n^{(2)*}) & \longrightarrow & \hat{\theta}^{(2)*}; \\ \quad ... \quad\quad\quad\quad\quad\quad & ... & \quad ... \\ \mathbf{x}^{(B)*} = (x_1^{(B)*}, ..., x_n^{(B)*}) & \longrightarrow & \hat{\theta}^{(B)*}. \end{array}$$

Como se vai verificar, a esperança do estimador $\hat{\theta}$ pode ser aproximada pela média amostral das réplicas *bootstrap*, ou seja,

$$\mathrm{E}[\hat{\theta}] \approx \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)*}.$$

Como as amostras *bootstrap* são seleccionadas aleatoriamente, as variáveis aleatórias correspondentes a $\hat{\theta}^{(b)*}, b = 1, ..., B$, são independentes. Por outro lado, dado que as amostras *bootstrap* têm todas o mesmo número de observações e que são todas retiradas de $\hat{F}$, as variáveis $\hat{\theta}^{(b)*}, b = 1, ..., B$, são identicamente distribuídas. Assim, pela lei dos grandes números, é possível verificar que

$$\frac{1}{B} \sum_{b=1}^{B} \hat{\theta}^{(b)*} \xrightarrow[B \to \infty]{P} \mathrm{E}^*[\hat{\theta}],$$

onde $\mathrm{E}^*[\hat{\theta}]$ representa o valor esperado do estimador $\hat{\theta}$ na distribuição da amostra observada, quando esta é interpretada como uma população. Ou seja, se se considerar todas as n^n amostras *bootstrap* provenientes da amostra $\mathbf{x}$ e todas as estimativas $(\hat{\theta}^{(1)*}, ..., \hat{\theta}^{(n^n)*})$ obtidas com cada uma dessas amostras, então

$$\mathrm{E}^*[\hat{\theta}] = \sum_{b=1}^{n^n} \hat{\theta}^{(b)*} \mathrm{P}^*[\hat{\theta} = \hat{\theta}^{(b)*}] = \frac{1}{n^n} \sum_{b=1}^{n^n} \hat{\theta}^{(b)*}.$$

Note-se que P^* representa a função de probabilidade do estimador associada à distribuição $\hat{F}$. Por isso, para qualquer b, $\mathrm{P}^*[\hat{\theta} = \hat{\theta}^{(b)*}] = \mathrm{P}[\text{ocorrer } \mathbf{x}^{(b)*}] = 1/n^n$, para $b = 1, ..., n^n$.

Por outro lado, na distribuição da população, $\mathrm{E}^*[\hat{\theta}]$ representa uma esperança condicional, uma vez que foi calculada condicionada à ocorrência da amostra original $(x_1, ..., x_n)$. Ao considerar a amostra teórica $(X_1, ..., X_n)$, $\mathrm{E}^*[\hat{\theta}]$ define uma média amostral de n^n variáveis aleatórias. Logo, quando $\mathrm{E}^*[\hat{\theta}] \xrightarrow[n \to \infty]{\mathcal{P}} \mathrm{E}[\hat{\theta}]$, fica garantido que a média

amostral de B réplicas *bootstrap* é um estimador consistente de $\mathrm{E}[\hat{\theta}]$ e, por isso, pode dar um valor aproximado dessa esperança.

$\Diamond$

O *bootstrap* também pode ser utilizado para aproximar outras características das estatísticas. No entanto, apesar de ser uma ferramenta bastante potente, o *bootstrap* não pode ser directamente utilizado em estruturas com dependência temporal ou espacial, uma vez que a reamostragem simples destrói a estrutura de dependência existente entre as variáveis aleatórias. Assim, várias propostas surgiram na literatura, com o objectivo de adaptar o *bootstrap* a estruturas de dependência temporal. Um exemplo dessas propostas é a metodologia *bootstrap* por blocos que seguidamente se vai abordar.

4.2 *Bootstrap* por blocos em estruturas temporais

Considere-se uma série temporal com distribuição dependente de um parâmetro θ e note-se por $\mathbf{x} = (x_1, x_2, ..., x_n)$ uma amostra observada. A ideia fundamental da metodologia *bootstrap* por blocos (divulgada, por exemplo, em Davison e Hinkley (1997)) pressupõe a estacionaridade forte da série temporal. Consiste em agrupar as componentes da amostra num conjunto de k blocos $\{B_1, B_2, ..., B_k\}$, onde cada bloco é formado por um número l_i de observações sucessivas, ou seja,

$$B_i = (x_j, x_{j+1}, ..., x_{j+l_i-1}), \text{ para } i = 1, ..., k.$$

A partir do conjunto de blocos $\{B_1, B_2, ..., B_k\}$, seleccionam-se aleatoriamente e com reposição k' blocos, os quais são justapostos para formarem a amostra *bootstrap* ($B_1^{(1)*}$, $B_2^{(1)*}, ..., B_{k'}^{(1)*}$) de dimensão n. Equivalentemente, a amostra *bootstrap* também pode ser representada através das observações, obtendo-se $\mathbf{x}^{(1)*} = (x_1^{(1)*}, x_2^{(1)*}, ..., x_n^{(1)*})$. A partir da amostra *bootstrap* calcula-se uma réplica da estatística que se pretende estudar, $\hat{\theta}^{(1)*} = T(\mathbf{x}^{(1)*})$. Repetindo este procedimento B vezes, obtém-se o conjunto das réplicas *bootstrap* $\{\hat{\theta}^{(b)*},\ b = 1, ..., B\}$, a partir das quais se estuda a estatística $\hat{\theta}$.

Considerar-se-á que os blocos têm um número fixo de observações l, embora, como se vê no parágrafo anterior, esse número possa ser aleatório.

Para formar os blocos consideram-se l observações sucessivas, uma vez que as observações mais próximas, são as mais correlacionadas e que, por isso, traduzem melhor a

estrutura de dependência. Assim, a estrutura de dependência é preservada dentro de cada um dos blocos, no processo de reamostragem.

A ideia anterior não merece qualquer contestação, mas o conjunto de blocos usado para fazer a reamostragem já não é tão consensual. Assim, existem várias versões de *bootstrap* por blocos – salientam-se os blocos contíguos, sugeridos por Carlstein (1992), os blocos sobrepostos de Künsch (1989) e os blocos circulares, propostos por Politis e Romano (1992).

O *bootstrap* com blocos contíguos obriga a que os blocos considerados no processo de reamostragem sejam disjuntos. Por isso, este método tem o inconveniente de originar poucos blocos para serem reamostrados. A Figura 4.2 mostra um esquema de *bootstrap* por blocos contíguos.

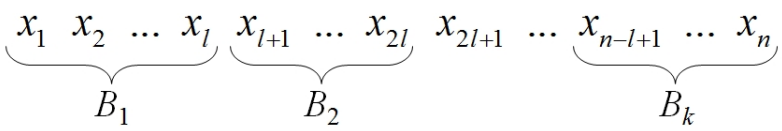

Figura 4.2: Ilustração da formação de blocos contíguos.

O *bootstrap* com blocos sobrepostos permite considerar mais blocos do que o *bootstrap* por blocos contíguos, uma vez que admite que os blocos se possam intersectar. No entanto, como se pode observar na Figura 4.3, as primeiras e as últimas observações da amostra integram menos blocos do que as do centro. Repare-se que cada uma das observações x_1 e x_n só fazem parte de um bloco, enquanto que as observações entre x_l e x_{n-l+1} já fazem parte de l blocos. Isto implica que as observações da amostra original não tenham a mesma probabilidade de ocorrência nas amostras *bootstrap*, pelo que o método conduz frequentemente a enviesamentos.

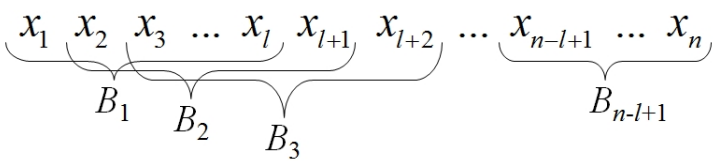

Figura 4.3: Ilustração da formação de blocos sobrepostos.

Para assegurar que todas as observações da amostra sejam reamostradas com a mesma probabilidade, Politis e Romano (1992) sugeriram que os extremos da estrutura temporal fossem ligados de modo a formarem uma cadeia circular, tal como se pode ver na Figura 4.4. Sendo assim, os blocos são formados a partir da ex-

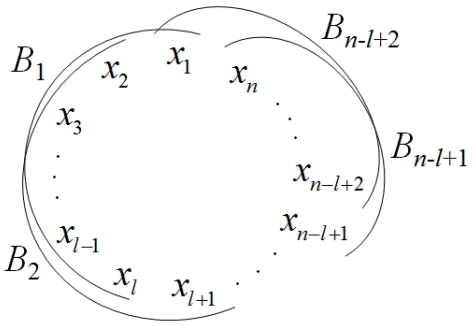

Figura 4.4: Ilustração da formação de blocos circulares.

tensão $(x_1, x_2, ..., x_n, x_{n+1}, ..., x_{n+l-1})$ da amostra original, do seguinte modo: para cada $i = 1, ..., l-1$, toma-se $x_{n+i} = x_i$. Obtém-se então um conjunto de n blocos $\{B_1, B_2, ..., B_n\}$, onde $B_i = (x_i, ..., x_{i+l-1})$, para $i = 1, ..., n$.

Em qualquer um dos métodos anteriormente referidos, depois de se formarem os blocos, seleccionam-se, aleatoriamente e com reposição, $k' = \frac{n}{l}$ blocos, para obter as amostras *bootstrap* $\mathbf{x}^{(b)\star} = (B_1^{(b)\star}, B_2^{(b)\star}, ..., B_{k'}^{(b)\star})$ e as correspondentes réplicas *bootstrap* $\hat{\theta}^{(b)\star} = T(\mathbf{x}^{(b)\star})$, para $b = 1, ..., B$. A escolha de $k' = \frac{n}{l}$ garante que as amostras *bootstrap* têm o mesmo número de variáveis aleatórias do que a amostra original.

Note-se que as metodologias *bootstrap* por blocos apresentadas pressupõem que as variáveis aleatórias podem trocar de posição nas amostras *bootstrap*. Para que essas permutações se possam fazer sem causarem problemas, é essencial que a série temporal apresente estacionaridade forte.

O *bootstrap* de blocos circulares é o que tem melhores propriedades – proporciona um número muito maior de blocos a reamostrar do que os blocos contíguos e, além disso, não causa o enviesamento já revelado que ocorre com os blocos sobrepostos. Assim, seguidamente, estudar-se-á a adaptação, a estruturas espaciais, da metodologia *bootstrap* por blocos circulares.

4.3 *Bootstrap* espacial por blocos circulares

Tal como acontece no contexto temporal, para adaptar o *bootstrap* a um contexto espacial é vantajosa a utilização de blocos. Os blocos permitem conservar a estrutura de dependência do processo durante o método de reamostragem. Hall (1985) constatou isso mesmo, quando estudou o método de reamostragem por blocos em estruturas espaciais. Contudo, ele analisou apenas os métodos espaciais análogos ao *bootstrap* por blocos contíguos e ao *bootstrap* por blocos sobrepostos (das séries temporais).

O procedimento de reamostragem que a seguir se sugere, segue de perto a fundamentação do *bootstrap* por blocos circulares das séries temporais, por ser aquele que apresenta as melhores propriedades. O método será denotado pela sigla *CMBB* em abreviatura de *Circular Moving Blocks Bootstrap*.

Uma vez que no espaço não existe a relação de ordem natural que existe no tempo, a construção dos blocos não pode ser baseada na noção de observações sucessivas, mas na de proximidade – como no espaço as observações mais correlacionadas são as que estão mais próximas, para que os blocos traduzam bem a estrutura de dependência, é essencial que sejam construídos à custa de observações próximas entre si. Por outro lado, para que as amostras *bootstrap* tenham a mesma disposição espacial da amostra original, é necessário que a forma e a dimensão dos blocos sejam tais que permitam pavimentar a amostra original.

Assim, considere-se um processo espacial univariado $\{Z(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^2\}$, que satisfaz a estacionaridade forte da **Definição 1.2.1**. Admita-se ainda que se dispõe de uma amostra do processo $Z(\mathbf{s})$ ao longo de uma grelha regular com n_y linhas e n_x colunas, ou seja, que se tem um conjunto de observações $\{X_{i,j} = Z((s_i, s_j)) : i = 1, ..., n_y \wedge j = 1, ..., n_x\}$, tal como mostra a Figura 4.5.

Em primeiro lugar, é necessário formar o conjunto dos blocos donde se faz a reamostragem. Fazendo uma analogia com o procedimento em estruturas temporais, os blocos são construídos a partir de uma extensão da amostra original, com $n_y + l_y - 1$ linhas e $n_x + l_x - 1$ colunas. A amostra ampliada pode ser construída do seguinte modo: para $i = 1, ..., n_y$ e $j = 1, ..., n_x$, as observações $x_{i,j}$ da amostra ampliada coincidem com a amostra original; se $i = n_y + 1, ..., n_y + l_y - 1$, ou se $j = n_x + 1, ..., n_x + l_x - 1$,

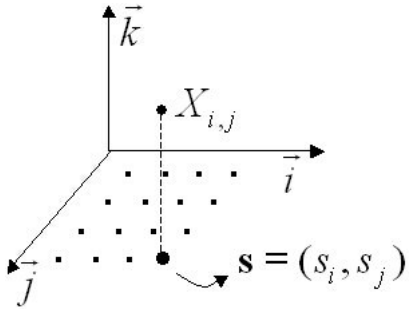

Figura 4.5: Observação $X_{i,j}$ do processo espacial, o qual foi amostrado ao longo de uma grelha regular.

as observações $x_{i,j}$ são iguais a $x_{i \bmod(n_y),j \bmod(n_x)}$ da amostra original, onde $a \bmod(b)$ representa o resto da divisão inteira de a por b. A Figura 4.6 procura ilustrar a amostra original e a ampliada. Note-se que, considerar a extensão da amostra equivale a transformar o plano das observações da amostra original num *torus* e, por isso, os blocos serão formados sobre esse *torus*.

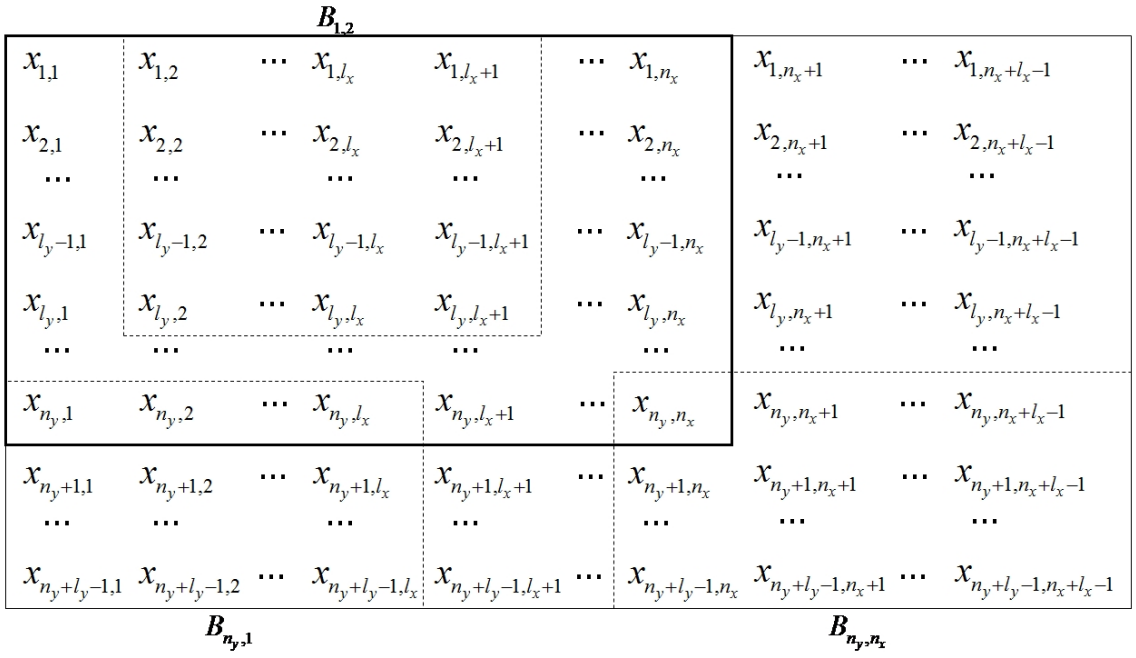

Figura 4.6: Extensão da amostra – indicação da mostra original, a cheio, e dos blocos $B_{1,2}$, $B_{n_y,1}$ e B_{n_y,n_x}, a tracejado.

Uma vez que se consideram observações em grelha, para que os blocos pavimentem

a amostra original e contenham observações próximas entre si, devem ser de forma rectangular, contendo l_y linhas e l_x colunas, com n_y divisível por l_y e n_x divisível por l_x. Deste modo, forma-se um conjunto de blocos que contém todos os blocos que se podem formar com l_y linhas e l_x colunas, a partir da amostra ampliada. Será desse conjunto de blocos que se faz a reamostragem. Por outras palavras, faz-se a reamostragem a partir do conjunto de blocos $\{B_{1,1}, B_{1,2}, ..., B_{n_y,n_x}\}$ (que contém $n_y \times n_x$ blocos), onde cada bloco é da forma

$$B_{p,q} = \{x_{i,j} : i = p, ..., p + l_y - 1 \wedge j = q, ..., q + l_x - 1\}, \tag{4.3.1}$$

para $p = 1, ..., n_y$ e $q = 1, ..., n_x$.

Para facilitar, identifica-se cada bloco pelo índice da observação do canto superior esquerdo, a qual está evidenciada na Figura 4.6. Este procedimento não implica a perda de generalidade, uma vez que existe uma bijecção entre o conjunto dos índices de localização das observações e o conjunto dos blocos.

Para construir a primeira amostra *bootstrap* $\mathbf{x}^{(1)*} = (B_{1,1}^{(1)*}, B_{1,2}^{(1)*}, ..., B_{k_y,k_x}^{(1)*})$, retiram-se $k_y \times k_x$ blocos, aleatoriamente e com reposição, do conjunto $\{B_{1,1}, ..., B_{n_y,n_x}\}$. Como se mostra na Figura 4.7, as amostras *bootstrap* são constituídas por uma grelha de k_y linhas e k_x colunas de blocos justapostos.

$B_{1,1}^{(b)*}$	$B_{1,2}^{(b)*}$	...	$B_{1,k_x}^{(b)*}$
$B_{2,1}^{(b)*}$	$B_{2,2}^{(b)*}$	...	$B_{2,k_x}^{(b)*}$
...	...	...	...
$B_{k_y,1}^{(b)*}$	$B_{k_y,2}^{(b)*}$	...	$B_{k_y,k_x}^{(b)*}$

Figura 4.7: A amostra *bootstrap* representada em termos dos blocos reamostrados.

Para que as amostras *bootstrap* tenham a mesma configuração da amostra original, é necessário que tenham n_y linhas e n_x colunas de observações. Isso acontece quando $k_y = \frac{n_y}{l_y}$ e $k_x = \frac{n_x}{l_x}$, o que define o número de blocos a reamostrar.

Uma vez concluída a formação dos blocos e tal como no *bootstrap* clássico, para cada amostra *bootstrap* gerada, calcula-se a correspondente réplica *bootstrap* $\hat{\theta}^{(b)*} = T(\mathbf{x}^{(b)*})$,

com $b = 1, ..., B$. Obtém-se, então, o conjunto das réplicas *bootstrap* $\{\hat{\theta}^{(1)*}, ..., \hat{\theta}^{(B)*}\}$, a partir do qual se faz o posterior estudo da estatística de interesse.

4.3.1 Estudo do enviesamento da média amostral

Para concretizar o esquema de reamostragem proposto e o seu interesse, considere-se a estimação de $\mathrm{E}[Z(\mathbf{s})] = \mu$ através da média amostral

$$\overline{X} = \sum_{i=1}^{n_y} \sum_{j=1}^{n_x} X_{i,j},$$

onde $X_{i,j}$, para $i = 1, ..., n_y$ e $j = 1, ..., n_x$, são as observações do processo $Z(\mathbf{s})$ (*vide* Figura 4.5). Quando se considera um processo geoestatístico com estacionaridade forte (*vide* **Definição 1.2.1**), sabe-se que $\overline{X}$ é centrada para estimar μ. Interessa agora verificar se a média amostral obtida com o *bootstrap* espacial proposto, notada por $\overline{X}^*$, goza de propriedades idênticas. Os resultados que se seguem asseguram que $\overline{X}^*$ converge em probabilidade para o valor observado $\overline{x}$.

Para o verificar, considere-se o estimador do enviesamento da média amostral, definido por

$$\hat{\beta}_B = \frac{1}{B} \sum_{b=1}^{B} \overline{X}^{(b)*} - \overline{x},$$

onde $\overline{X}^{(b)*}$ e $\overline{x}$ representam, respectivamente, a média amostral da b-ésima amostra *bootstrap* e o valor da média amostral na amostra real. As variáveis $\overline{X}^{(b)*}$, $b = 1, ..., B$, são *i.i.d.*, com esperança e variância finita. Logo, pela lei dos grandes números, pode-se concluir que

$$\frac{1}{B} \sum_{b=1}^{B} \overline{X}^{(b)*} \xrightarrow[B \to \infty]{P} \mathrm{E}^*[\overline{X}^*],$$

onde $\mathrm{E}^*[\overline{X}^*]$ denota a esperança da média amostral na distribuição $\hat{F}$.

Resta agora averiguar qual é o valor de $\mathrm{E}^*[\overline{X}^*]$. Para tal, para qualquer $p = 1, ..., n_y$ e $q = 1, ..., n_x$, denote-se por $S^*_{p,q}$ a soma de todas as variáveis aleatórias contidas no bloco $B^*_{p,q}$, onde $B^*_{p,q}$ representa um bloco reamostrado da forma (4.3.1); assim,

$$S^*_{p,q} = \sum_{i=p}^{p+l_y-1} \sum_{j=q}^{q+l_x-1} X^*_{i,j}.$$

É possível escrever $\mathrm{E}^*[\overline{X}^*]$ como função dos $S^*_{p,q}$, obtendo-se

$$\mathrm{E}^*[\overline{X}^*] = \frac{1}{n_x n_y}\left[\mathrm{E}^*[S^*_{1,1}] + \mathrm{E}^*[S^*_{1,l_x+1}] + ... + \mathrm{E}^*[S^*_{n_y-l_y+1,n_x-l_x+1}]\right]. \tag{4.3.2}$$

Para calcular cada uma das parcelas em (4.3.2) note-se que, para p' e q' fixos (com $p' = 1, l_y+1, 2l_y+1, ..., n_y-l_y+1$ e $q' = 1, l_x+1, 2l_x+1, ..., n_x-l_x+1$) se tem que

$$\mathrm{E}^*[S^*_{p',q'}] = \sum_{s_{p,q}\in\mathcal{S}} s_{p,q}\mathrm{P}[S^*_{p',q'} = s_{p,q}],$$

onde $\mathcal{S}$ é o conjunto das somas que se obtêm quando, em cada um dos blocos que podem ser reamostrados, se somam todos os $x_{i,j}$, ou seja,

$$\mathcal{S} = \left\{ s_{p,q} : s_{p,q} = \sum_{i=p}^{p+l_y-1}\sum_{j=q}^{q+l_x-1} x_{i,j},\ p = 1, ..., n_y \wedge q = 1, ..., n_x \right\}.$$

Como os blocos são seleccionados aleatoriamente e com reposição, todos os elementos do suporte têm a mesma probabilidade $\frac{1}{n_x n_y}$ de ocorrer. Assim,

$$\mathrm{E}^*[S^*_{p',q'}] = \frac{1}{n_x n_y}\sum_{p=1}^{n_y}\sum_{q=1}^{n_x} s_{p,q} = \frac{1}{n_x n_y}\sum_{p=1}^{n_y}\sum_{q=1}^{n_x}\sum_{i=p}^{p+l_y-1}\sum_{j=q}^{q+l_x-1} x_{i,j}.$$

Ao serem considerados todos os blocos, cada $x_{i,j}$ vai ocorrer $l_x l_y$ vezes, logo

$$\mathrm{E}^*[S^*_{p',q'}] = \frac{1}{n_x n_y}\sum_{i=1}^{n_y}\sum_{j=1}^{n_x} l_x l_y x_{i,j} = l_x l_y \overline{x}.$$

Então, substituindo cada $\mathrm{E}^*[S^*_{p',q'}]$ pelo seu valor na equação (4.3.2) e tendo em conta que existem $k_x k_y$ parcelas, vem que

$$\mathrm{E}^*[\overline{X}^*] = \frac{1}{n_x n_y}\left[k_x k_y \times l_x l_y \overline{x}\right] = \overline{x}.$$

Daqui resulta que, $\frac{1}{B}\sum_{b=1}^{B}\overline{X}^{(b)*} \xrightarrow[B\to\infty]{P} \overline{x}$. Dado que o valor $\overline{x}$ foi observado e, portanto, não depende do número de réplicas B, conclui-se que

$$\hat{\beta}_B = \frac{1}{B}\sum_{b=1}^{B}\overline{X}^{(b)*} - \overline{x} \xrightarrow[B\to\infty]{P} 0.$$

Portanto, com o esquema de reamostragem proposto, à medida que o número de réplicas aumenta, o estimador *bootstrap* do enviesamento da média amostral tende em probabilidade para o valor correcto, *i.e*, tende em probabilidade para zero.

4.3.2 Exemplo de simulação

Para ilustrar o comportamento da metodologia *bootstrap* proposta, apresenta-se um estudo de simulação, elaborado com o programa *R*, usando a *package geoR* que permite trabalhar com os processos geoestatísticos. Com a função *grf*, que simula processos geoestatísticos, geraram-se processos espaciais Gaussianos de variograma isotrópico. Utilizou-se a configuração de base da função *grf* e introduziram-se os seguintes parâmetros: média do processo nula; variograma esférico (*vide* ponto (1.3.7)) de amplitude $\phi = 5$, patamar $\sigma^2 = 2$ e efeito de pepita τ^2 nulo.

Consideraram-se amostras constituídas por observações dispostas ao longo de grelhas quadrangulares de 12, 18, 24 e 30 observações de lado. Para cada grelha amostral, observaram-se 500 processos gerados nas condições referidas. Usando cada uma das amostras geradas e fixando a dimensão do bloco, utilizou-se o *CMBB*, primeiro com 200 réplicas e depois com 1000 réplicas *bootstrap*. De seguida, foram considerados blocos de diferentes dimensões. Em cada caso, estimou-se o enviesamento da média amostral com o *CMBB*.

Apenas se apresenta detalhadamente os resultados obtidos com a grelha amostral de 24 observações de lado, uma vez que os restantes casos conduziram a idênticas conclusões. Para estudar o efeito da dimensão do bloco em relação à dimensão da amostra e à amplitude do processo, foram considerados blocos quadrangulares de 2, 3, 4, 6, 8 e 12 observações de lado.

Na Figura 4.8 apresentam-se os diagramas de extremos e quartis das estimativas *CMBB* obtidas para o enviesamento da média amostral, considerando os blocos de diferentes dimensões e para cada um dos números de réplicas estudados. As rectas horizontais a tracejado representam o limite em probabilidade do enviesamento que, como se viu na subsecção **4.3.1**, é zero.

Observando a Figura 4.8, verifica-se que a mediana das estimativas é muito próxima de zero e que, quando o número de réplicas aumenta, a dispersão das estimativas diminui.

Em cada um dos casos estudados (200 e 1000 réplicas *bootstrap*), calculou-se ainda

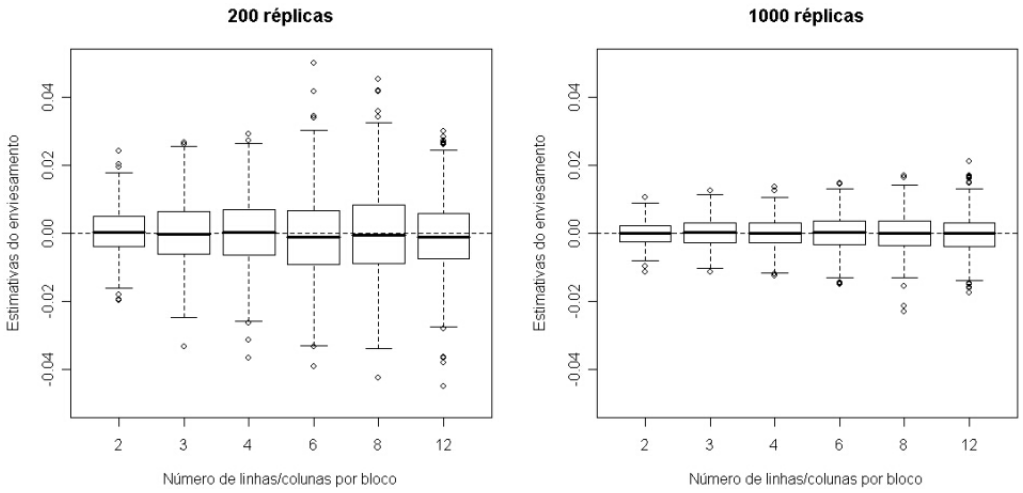

Figura 4.8: Diagramas de extremos e quartis das estimativas *CMBB* do enviesamento da média amostral, para uma grelha amostral com 24 observações de lado.

os erros quadráticos médios empíricos relativos aos 500 processos, através de

$$EQME(\hat{\beta}_B) = \frac{1}{500}\sum_{i=1}^{500}[\hat{v}_i - v]^2,$$

onde $\hat{v}_i, i = 1, ..., 500$, designa a i-ésima estimativa *bootstrap* do enviesamento da média amostral e $v = 0$ é o valor teórico do enviesamento. Os respectivos resultados encontram-se na Tabela 4.1.

$l_x = l_y$	$B = 200$	$B = 1000$
2	0.46×10^{-4}	0.10×10^{-4}
3	0.86×10^{-4}	0.16×10^{-4}
4	1.04×10^{-4}	0.21×10^{-4}
6	1.57×10^{-4}	0.27×10^{-4}
8	1.71×10^{-4}	0.31×10^{-4}
12	1.28×10^{-4}	0.30×10^{-4}

Tabela 4.1: Erros quadráticos médios empíricos do estimador *CMBB* do enviesamento da média amostral, para a grelha com 24 observações de lado, variando o comprimento do lado dos blocos.

Observando a tabela verifica-se que, em geral, os erros quadráticos médios empíricos tendem a aumentar com o comprimento do lado do bloco, mas que são sempre extremamente reduzidos. Além disso, pode-se ainda confirmar que as estimativas do enviesamento da média amostral melhoram quando se aumenta de 200 para 1000 réplicas.

Este exemplo confirma que, nas situações consideradas, a metodologia *CMBB* dá bons resultados.

Capítulo 5

Estudo da estacionaridade da média do processo

Desde que exista a média de um processo geoestatístico, ao admitir que ele é estacionário, fica implícita a condição de que o processo tem média constante. Em geral, a estacionaridade da média é assumida como verdadeira; no entanto, não são conhecidos métodos objectivos que permitam testar essa hipótese. Esse facto deve-se, sobretudo, a dois tipos de dificuldades: por um lado, ao facto de, em cada localização, existir apenas uma realização de uma única variável aleatória, o que dificulta a construção de testes para o valor médio; por outro lado, as condições de dependência das variáveis observáveis tornam desconhecidas as distribuições das estatísticas usuais.

Neste capítulo propõe-se um teste estatístico para confirmar a estacionaridade da média. O teste recorre a um método de projecções e é desenvolvido num contexto de regressão com erros dependentes. Procurou-se usar um estimador robusto dos coeficientes de regressão como estatística do teste, aproximando a sua distribuição através de um método *bootstrap* que não exige condições de estacionaridade forte. Assim, ao contrário do método focado no capítulo anterior, a versão que aqui se apresenta é aplicável a processos estacionários de segunda ordem.

Sendo assim, começa-se por estudar a distribuição assintótica do estimador de mínimos desvios absolutos (LAD) e dos estimadores-MM, num modelo de regressão linear onde os erros apresentam uma estrutura de dependência específica. Nas restantes secções do capítulo faz-se a apresentação detalhada da proposta do teste. Embora o teste possa ser desenvolvido com qualquer uma das duas estatísticas, quando o processo tem variância constante, recomenda-se a utilização dos estimadores-MM, devido

às suas boas propriedades.

5.1 Distribuição assintótica do LAD sob observações m-dependentes

Considere-se o modelo definido por

$$Z_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^k + \epsilon_i, \quad i = 1, ..., n,$$

tal que

$$\mathrm{E}[Z_i|x_i] = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^k, \quad p \in \mathbb{N}, \tag{5.1.1}$$

onde $(\beta_0, ..., \beta_p)$ são parâmetros desconhecidos, x_i são as variáveis explicativas (valores fixos), Z_i são variáveis aleatórias contínuas e ϵ_i são os erros do modelo, os quais têm média nula e variância constante $(i = 1, ..., n)$. O que distingue este modelo do modelo tradicional de regressão linear é o facto dos erros ϵ_i serem dependentes. Neste estudo, supor-se-á que os erros apresentam uma forma determinada de dependência, em particular, que são m-dependentes. Este conceito, que se passa a apresentar, foi utilizado, por exemplo, por Christofides e Mavrikiou (2003) considerando a norma euclidiana.

Definição 5.1.1. Seja $X(\mathbf{s}), \mathbf{s} \in D$, um processo geoestatístico e $m \in \mathbb{R}_0^+$ uma constante. O processo $X(\mathbf{s})$ diz-se *m-dependente* se, para quaisquer $\mathbf{s}_i$ e $\mathbf{s}_j$ do domínio D,

$$\|\mathbf{s}_i - \mathbf{s}_j\| > m \Rightarrow \mathrm{Cov}[X(\mathbf{s}_i), X(\mathbf{s}_j)] = 0,$$

onde $\|.\|$ denota uma norma.

$\Diamond$

Considere-se o caso particular do modelo (5.1.1) com erros m-dependentes, o que significa que, sempre que $\|x_i - x_j\| > m$ então $\mathrm{Cov}[\epsilon_i, \epsilon_j] = 0$. Neste caso, $\|.\|$ representa a norma euclidiana. O facto dos erros serem m-dependentes implica, imediatamente, que as observações Z_i também o sejam.

Quando os erros do modelo têm distribuição simétrica, a mediana e a média coincidem, o que permite testar o valor da média construindo um teste para o valor da

mediana. De acordo com esta via, o modelo de regressão quantílica é especialmente interessante, uma vez que não impõe a homoscedasticidade. Note-se que um processo $Z(\mathbf{s})$ com estacionaridade de segunda ordem tem variância constante; mas se $Z(\mathbf{s})$ for apenas intrinsecamente estacionário, essa condição não está assegurada.

O modelo de regressão quantílica foi sugerido por Koenker e Basset (1978). A sua principal vantagem em relação ao modelo de regressão tradicional, reside no facto da regressão quantílica permitir estudar a distribuição condicional em diferentes quantis, ao contrário da regressão tradicional que apenas considera a média condicional.

Dadas as observações $(\mathbf{x}_i, Z_i)$, para $i = 1, ..., n$, onde $\mathbf{x}_i$ é o vector das variáveis explicativas, o modelo linear de regressão quantílica caracteriza-se por

$$Z_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\theta + \epsilon_{\theta,i},$$

onde os pares observados verificam a equação quantílica

$$Q_\theta(Z_i|\mathbf{x}_i) = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_\theta,$$

para $\theta \in]0, 1[$ fixo. As variáveis $\epsilon_{\theta,i}$ são os erros aleatórios do modelo, $Q_\theta(Z_i|\mathbf{x}_i)$ denota o quantil de ordem θ da distribuição condicional $Z_i|\mathbf{x}_i$ e $\boldsymbol{\beta}_\theta \in \mathbb{R}^{p+1}$ é o vector dos parâmetros desconhecidos que se pretende estimar.

O estimador de regressão quantílica de $\boldsymbol{\beta}_\theta$ é implicitamente definido pela equação

$$\hat{\boldsymbol{\beta}}_{RQ} = \arg\min_{\boldsymbol{\beta}_\theta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_{\theta,i}(\theta - \mathbb{I}_{]-\infty,0[}(\epsilon_{\theta,i})), \tag{5.1.2}$$

onde $\mathbb{I}_A$ representa a função indicatriz do intervalo A. Os estimadores $\hat{\boldsymbol{\beta}}_{RQ}$ são robustos no sentido infinitesimal, uma vez que têm função de influência limitada. Para além disso, têm um ponto de ruptura positivo que depende do quantil θ. Repare-se ainda que estes estimadores pertencem à família dos estimadores-M já mencionados na subsecção **3.3.3**.

Admitindo que as observações Z_i têm distribuições simétricas, então elas têm a mediana igual à média. Consequentemente, para formalizar o modelo de regressão quantílica com $Q_{0.5}(\epsilon_i|x_i) = \text{Mediana}(\epsilon_i|x_i) = 0$, a equação (5.1.1) é substituída por uma equação da forma

$$\mathrm{E}[Z_i|x_i] = \text{Mediana}[Z_i|x_i] = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^k. \tag{5.1.3}$$

Assim, tomando $\theta = 50\%$ em (5.1.2), obtém-se o estimador de mínimos desvios absolutos (*Least Absolute Deviation estimator*, abreviado por LAD). O estimador LAD para o modelo (5.1.3) é definido por

$$\hat{\boldsymbol{\beta}}_{LAD} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left| Z_i - \beta_0 - \sum_{k=1}^{p} \beta_k x_i^k \right|. \tag{5.1.4}$$

Nesta secção pretende-se conhecer a distribuição assintótica do LAD, supondo que os erros do modelo são m-dependentes.

A distribuição assintótica do estimador LAD é bem conhecida num contexto de observações $i.i.d.$. Nesse caso, Koenker e Basset (1978) mostraram que, sob condições de regularidade, o LAD tem distribuição assintótica normal. Contudo, a distribuição assintótica não é conhecida no caso geral de erros dependentes. Existem alguns resultados para formas específicas de dependência, tais como modelos ARMA e ARCH (ver, por exemplo, em Koenker (2005)). Também existem alguns resultados em condições de dependência que são baseados na metodologia *bootstrap*. Neste campo salienta-se o trabalho de Fitzenberger (1997). Fitzenberger estudou a distribuição do estimador de mínimos quadrados e dos estimadores de regressão quantílica, e provou que todos eles possuem distribuição assintótica normal, considerando que os erros do modelo verificam condições de dependência forte (*strong mixing conditions*). Essas condições, tal como o próprio nome indica, permitem que a dependência entre as variáveis aleatórias, possa ser mais forte do que nas condições de m-dependência (*vide* Bradley (2005)). Sendo assim, os processos com erros m-dependentes constituem um subconjunto dos processos que verificam condições de dependência forte, por isso, gozam de propriedades características da última família referida. O facto da m-dependência implicar a dependência forte vai ser utilizado na demonstração da **Proposição 5.1.1**, para encontrar a distribuição assintótica do LAD em modelos com erros m-dependentes. O resultado é válido sob condições idênticas às que foram assumidas em Fitzenberger (1997) para a dependência forte.

Proposição 5.1.1. Sejam $\{(x_1, Z_1), (x_2, Z_2), ..., (x_n, Z_n)\}$ observações do modelo de regressão linear (5.1.3), onde x_i são os regressores fixos do modelo e Z_i são variáveis aleatórias contínuas, simétricas e m-dependentes ($i = 1, ..., n$). Considere-se o estimador da regressão quantílica com $\theta = 50\%$, definido em (5.1.4), e denote-se por $\hat{\boldsymbol{\beta}}_{LAD} =$

$(\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$ o estimador LAD de $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$. Seja $\mathbf{D}_n$ a matriz diagonal definida por

$$\mathbf{D}_n = diag\left(1, \frac{1}{n}\sum_{i=1}^{n} x_i^2, \frac{1}{n}\sum_{i=1}^{n}(x_i^2)^2, ..., \frac{1}{n}\sum_{i=1}^{n}(x_i^p)^2\right),$$

e d_1 e d_2 duas constantes tais que

C.1 $\max\limits_{1\leq i\leq n} \mathbf{x}_i^{\mathrm{T}}\mathbf{D}_n^{-1}\mathbf{x}_i < d_1$, onde $\mathbf{x}_i$ são os vectores coluna $[1\ x_i\ ...\ x_i^p]^{\mathrm{T}}$;

C.2 a matriz $d_2\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}} - n\mathbf{D}_n$ é definida positiva para qualquer $n \in \mathbb{N}$, no sentido em que todos os seus valores próprios são positivos; onde $\mathbf{X}$ é a matriz $[\mathbf{x}_1\ ...\ \mathbf{x}_n]$, $\boldsymbol{\Sigma}$ representa a matriz de covariâncias do processo $\{sgn(\epsilon_i)\}$ e sgn representa a função sinal usual.

Considere-se ainda que a matriz $n^{-1}\mathbf{D}_n^{-1/2}\mathbf{C}_n$ é uniformemente definida positiva, onde $\mathbf{C}_n$ é igual a $\sum_{i=1}^{n} f_i(0)\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}$ e f_i é a função densidade de ϵ_i.

Se se verificarem as condições referidas anteriormente, então

$$\hat{\boldsymbol{\beta}}_{LAD} \xrightarrow{\mathcal{P}} \boldsymbol{\beta},$$

onde $\mathcal{P}$ significa convergência em probabilidade, e

$$\hat{\boldsymbol{\beta}}_{LAD} \xrightarrow{\mathcal{L}} N_{p+1}\left(\boldsymbol{\beta}, \mathbf{V}_n\right),$$

onde $\mathcal{L}$ representa a convergência em lei e N_{p+1} designa a distribuição normal multivariada de dimensão $p+1$ que, neste caso, tem matriz de covariâncias $\mathbf{V}_n$ que é igual a $\mathbf{C}_n^{-1}\mathbf{X}\boldsymbol{\Sigma}\mathbf{X}^{\mathrm{T}}\mathbf{C}_n^{-1}$.

Demonstração: Em Bradley (2005) é possível verificar que, se um conjunto de variáveis aleatórias é m-dependente, então esse conjunto verifica as condições de dependência forte. Assim, dado que as observações $\{Z_1, Z_2, ..., Z_n\}$ são m-dependentes, então elas apresentam condições de dependência forte e, consequentemente, os erros $\epsilon_i, i = 1, ..., n$, do modelo de regressão também.

Por outro lado, dado que as variáveis aleatórias do processo $Z(\mathbf{s})$ têm distribuição simétrica, então os erros ϵ_i do modelo de regressão também têm distribuição simétrica. Assim

$$\mathrm{E}[\epsilon_i] = \mathrm{Mediana}[\epsilon_i] = 0.$$

Isto implica que

$$\mathrm{E}[sgn(\epsilon_i)] = \int_{-\infty}^{0} -f_i(u)\, du + \int_{0}^{\infty} f_i(u)\, du = 0$$

para todo o $i = 1, ..., n$.

Uma vez que, por hipótese,

- se verificam as condições **C.1** e **C.2**;
- a matriz $n^{-1}\mathbf{D}_n^{-1/2}\mathbf{C}_n$ é uniformemente definida positiva,

e que

- os erros do modelo verificam as condições de dependência forte;
- $\mathrm{E}[sgn(\epsilon_i)] = 0$ para qualquer $i = 1, ..., n$,

então o Teorema C.2 de Fitzenberger (1997) permite concluir o resultado pretendido.

□

Note-se que a forma da matriz de covariâncias $\mathbf{V}_n$ da proposição anterior é conhecida, mas as funções densidade f_i dos erros do modelo, bem como a matriz $\mathbf{\Sigma}$, não. Por isso, é impossível estimar a matriz $\mathbf{V}_n$ pelos métodos tradicionais de inferência. Torna-se então necessário recorrer à metodologia *bootstrap* para aproximar a estrutura de covariâncias do estimador *LAD*. Este tema voltará a ser abordado com mais detalhe na subsecção **5.3.2**.

5.2 Distribuição assintótica dos estimadores-MM sob observações m-dependentes

Considere-se o modelo de regressão linear apresentado em (5.1.1). Pela **Definição 3.3.10**, o estimador-MM de $\boldsymbol{\beta}$, que se vai denotar por $\hat{\boldsymbol{\beta}}_{MM}$, é o conjunto de todas as soluções do estimador-M de equação

$$\sum_{i=1}^{n} \rho_1' \left(\frac{Z_i - \mathbf{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}}_{MM}}{\hat{\sigma}_n} \right) \mathbf{x}_i = 0,$$

onde ρ_1' representa a derivada da função ρ_1 e $\hat{\sigma}_n$ é a estimativa do estimador-S de escala, a qual é o valor mínimo de $\hat{\sigma}_n(\boldsymbol{\beta})$ (com $\boldsymbol{\beta}$ a variar em $\mathbb{R}^{p+1}$) que satisfaz a equação

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{Z_i-\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}}{\hat{\sigma}_n(\boldsymbol{\beta})}\right)=b,$$

onde b é a constante de afinação.

Quando os regressores são fixos e os erros do modelo são *i.i.d.*, Salibian-Barrera (2006) mostrou que os estimadores-MM são fortemente consistentes para $\boldsymbol{\beta}$ e que têm uma distribuição assintótica normal. Contudo, a distribuição assintótica destes estimadores é desconhecida em situações onde existe uma estrutura de dependência nos erros do modelo. Por isso, nesta secção apresenta-se uma metodologia que permite estudar a distribuição assintótica dos estimadores-MM, quando os erros do modelo são m-dependentes (*vide* **Definição 5.1.1**).

A abordagem que aqui vai ser feita assume um modelo mais geral do que o modelo apresentado em (5.1.1). Neste caso estuda-se um processo espacial $Z(\mathbf{s})$ m-dependente, com domínio contido em $\mathbb{R}^2$, cuja amostra $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n)\}$ se encontra disposta ao longo de uma grelha regular do plano, tal como mostra a Figura 5.1. Assume-se

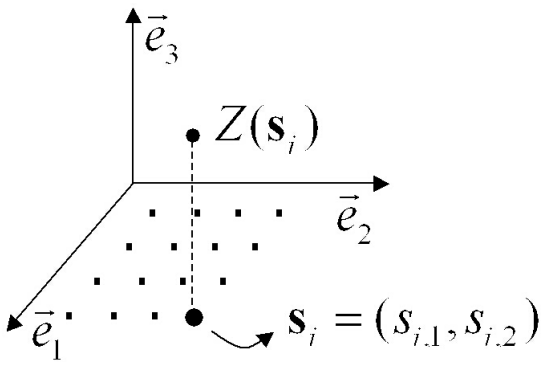

Figura 5.1: Representação das observações do processo $Z(\mathbf{s})$, as quais estão dispostas ao longo de uma grelha regular.

ainda que o processo $Z(\mathbf{s})$ pode ser representado através do modelo de regressão linear

$$Z(\mathbf{s}_i)=\sum_{j=0}^{p}\beta_j f_j(\mathbf{s}_i)+\delta(\mathbf{s}_i),\quad i=1,...,n,\quad p\in\mathbb{N},$$

com $\mathrm{E}[Z(\mathbf{s})|\mathbf{s}]=\sum_{j=0}^{p}\beta_j f_j(\mathbf{s})$, onde $\delta(\mathbf{s}_i)$ são os erros do modelo, que se supõe serem m-dependentes, e f_j são funções que só dependem da localização $\mathbf{s}$.

Determinar a distribuição assintótica dos estimadores-MM de $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ não é uma tarefa fácil, por causa da estrutura de dependência presente nos erros do modelo. Contudo, a distribuição assintótica dos estimadores-MM, pode ser aproximada através de uma forma específica de reamostragem por *bootstrap* espacial. Note-se que as metodologias *bootstrap* apresentadas no **Capítulo 4** não são adequadas à resolução deste problema – como esses métodos permitem a troca da posição relativa dos blocos durante o processo de reamostragem, a relação que existe entre a variável dependente e os regressores do modelo é destruída quando se permutam as observações. Consequentemente, é necessário considerar uma versão *bootstrap* que não altere a posição relativa entre os blocos, durante o processo de reamostragem. Por isso, no seguimento, utiliza-se um procedimento de reamostragem que é semelhante ao proposto por Lahiri (2003). Nos próximos parágrafos, descreve-se o método *bootstrap* utilizado.

Tal como acontece com os procedimentos *bootstrap* para observações dependentes no tempo, a metodologia *bootstrap* espacial que a seguir se considera, utiliza blocos para aproximar a distribuição assintótica dos estimadores-MM. Os blocos preservam no seu interior a estrutura de dependência do modelo. Neste caso, para que a estrutura de m-dependência dos erros do modelo seja reflectida dentro de cada um dos blocos, estes devem ser formados por observações próximas entre si. De preferência, devem ser formados por observações, cujas localizações estejam separadas por uma distância euclidiana inferior a m, visto que m é a distância a partir da qual as observações se podem considerar não correlacionadas.

Através da **Definição 5.1.1** (com a norma euclidiana) é possível concluir que, dado um ponto $\mathbf{s}_j \in D$, o conjunto de localizações $\mathbf{s}$ tais que as observações $Z(\mathbf{s}_j)$ e $Z(\mathbf{s})$ podem ser correlacionadas, é definido pelo interior de uma circunferência de centro em $\mathbf{s}_j$ e de raio m. Então, poder-se-ia concluir que os blocos ideais teriam a forma circular. No entanto, os blocos de forma circular não se ajustam bem à disposição das localizações. Como as localizações estão dispostas ao longo de uma grelha regular do plano, os blocos de forma circular vão desprezar as localizações que se encontram nos cantos da grelha e não vão reamostrar essas observações. Por isso, vão-se considerar blocos de forma quadrangular, para ser possível pavimentar a grelha da amostra original. Portanto, para formar os blocos, propõe-se que a distância euclidiana

seja substituída por outra distância d' tal que

$$\forall_{\mathbf{s}_i,\mathbf{s}_j \in D} \quad d'(\mathbf{s}_i, \mathbf{s}_j) = \max_{k=1,2} |s_{i,k} - s_{j,k}|. \tag{5.2.1}$$

Como consequência, os blocos passam a ser conjuntos da forma

$$b_j = \{(\mathbf{s}_i, Z(\mathbf{s}_i)) : d'(\mathbf{s}_i, \mathbf{s}_j) \leq L, 0 < L \leq m\},$$

que contêm as observações cujas localizações estão no interior de quadrados centrados em $\mathbf{s}_j, j = 1, ..., n$, e de lado $C = 2L$, tal como mostra a Figura 5.2.

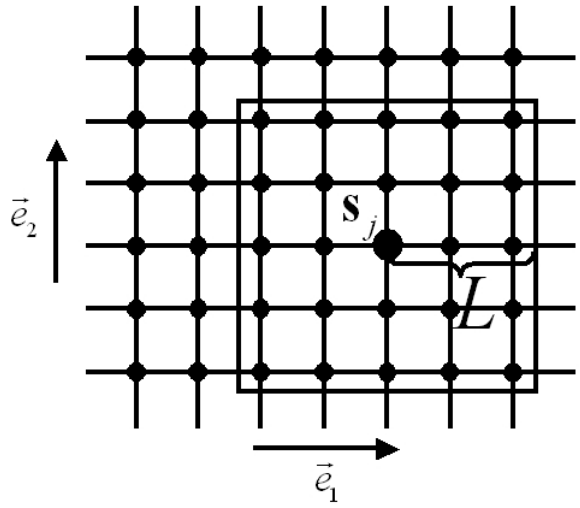

Figura 5.2: Ilustração dos blocos *bootstrap* formados usando a distância em (5.2.1) – cada localização $\mathbf{s}_j$ traz consigo os pontos situados numa vizinhança de raio L.

Ao contrário do que acontece com a versão *bootstrap CMBB* que foi proposta na secção **4.3**, nesta secção admite-se que a localização dos blocos reamostrados não se altera na amostra *bootstrap*. A reamostragem é conseguida obtendo blocos que podem estar sobrepostos. Isto permite que a posição relativa entre as observações se mantenha nas amostras *bootstrap* e, por isso, a relação que existe entre a variável dependente e os regressores do modelo, é preservada.

Supor-se-á que todos os blocos contêm o mesmo número de observações, designado por n_o. O número de blocos a reamostrar, n_b, deve ser tal que, em cada amostra *bootstrap*, se tenham tantas observações como na amostra original, isto é, as amostras *bootstrap* devem ter n observações. Portanto, o ideal seria reamostrar-se n_b blocos,

onde $n = n_b \times n_o$. No entanto, existem situações onde n não é um múltiplo de n_o, por isso, quando tal acontece, determina-se n_b como sendo o menor número inteiro que satisfaz a condição $n_b \times n_o \geq n$. Desse modo, fica assegurado que em cada amostra *bootstrap* existem, pelo menos, tantas observações quantas as que existem na amostra original.

Este método de reamostragem vai ser designado por *SFBB* (do Inglês *Spatial Fixed Blocks Bootstrap*) e pode ser resumido no seguinte algoritmo.

Algoritmo para o *SFBB*:

1. Escolhe-se, ao acaso, uma das localizações contidas no conjunto $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ da grelha da amostra original e denomina-se essa localização por $\mathbf{s}_j^*$.

2. Com a localização $\mathbf{s}_j^*$ forma-se o bloco $b_j^* = \{(\mathbf{s}_i, Z(\mathbf{s}_i)) : d'(\mathbf{s}_i, \mathbf{s}_j^*) \leq L, 0 < L \leq m\}$, onde $d'(\mathbf{s}_i, \mathbf{s}_j) = \max_{k=1,2} |s_{i,k} - s_{j,k}|$.

3. Para fixar o número de observações por bloco, consideram-se apenas blocos com $n_o = \max_{j=1,...,n_b}\{\#b_j^*\}$ observações. Caso a observação seleccionada dê origem a um bloco com menos observações do que n_o, ignora-se essa observação e selecciona-se outra, ao acaso.

4. Consideram-se sucessivos blocos b_j^*, até se obter a amostra *bootstrap* $(b_1^*, ..., b_{n_b}^*)$. O número total de blocos n_b é o menor inteiro que satisfaz a condição $n_b \times n_o \geq n$.

5. A partir da amostra *bootstrap* obtém-se uma réplica *bootstrap* $\hat{\beta}_k^{*(b)}$ do estimador-MM $\hat{\beta}_k$, para $k = 0, ..., p$.

6. Repetem-se os passos anteriores B vezes obtendo-se um conjunto de réplicas *bootstrap* $\left(\hat{\beta}_k^{*(1)}, ..., \hat{\beta}_k^{*(B)}\right)$ do estimador-MM de β_k, para $k = 0, ..., p$.

O conjunto das réplicas *bootstrap* $\left(\hat{\beta}_k^{*(1)}, ..., \hat{\beta}_k^{*(B)}\right)$ que se obtém através deste procedimento permite aproximar a distribuição dos estimadores-MM de β_k, para cada $k = 0, ..., p$, nas condições assumidas de erros m-dependentes.

Para ilustrar esta metodologia, seguidamente apresenta-se um exemplo de simulação, no qual se aproxima a distribuição assintótica dos estimadores-MM, num contexto de processos geoestatísticos.

Exemplo 5.2.1. Com a *package geoR* do programa R, geraram-se observações do processo geoestatístico

$$Z(s_1, s_2) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \delta(s_1, s_2), \quad \mathbf{s} = (s_1, s_2) \in D \subset \mathbb{R}^2,$$

onde os erros $\delta(s_1, s_2)$ do modelo foram gerados por um processo Gaussiano de média nula e de amplitude m igual a 5 unidades. Portanto, o processo $Z(\mathbf{s})$ é m-dependente com $m = 5$. Simularam-se amostras do processo ao longo de grelhas regulares do plano e, para averiguar como varia a distribuição dos estimadores-MM à medida que a dimensão da grelha aumenta, geraram-se amostras em grelhas de 10×10, 20×20, 30×30, 40×40, 50×50 e 60×60 localizações. Deste modo, analisou-se o comportamento assintótico do estimador de acordo com a metodologia *IDA* (*vide* subsecção **2.3.1**). Para cada dimensão da grelha geraram-se 100 amostras do processo $Z(\mathbf{s})$.

Para aproximar a distribuição dos estimadores-MM dos coeficientes ($\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\beta}_2$), utilizou-se o método *SFBB* em cada uma das amostras geradas. Para que o bloco fosse representativo da dependência entre observações, consideraram-se blocos de lado igual à distância m a partir da qual deixa de existir correlação. Assim, como se pode ver pela Figura 5.2, para um valor de m igual a 5 unidades, tomou-se $L = 2.5$. As estimativas obtidas com os estimadores-MM foram calculadas através da *package roblm* do programa R, com as opções predefinidas.

Uma análise preliminar dos resultados, sugeriu que a distribuição das réplicas *bootstrap* obtidas fosse aproximada pela distribuição normal. Essa constatação fez com que se efectuasse um teste de ajustamento à distribuição normal das réplicas *bootstrap*. Ou seja, com cada uma das amostras geradas e para cada k fixo, testou-se

$H_0 : \left(\hat{\beta}_k^{*(1)}, ..., \hat{\beta}_k^{*(B)}\right)$ provêm de uma distribuição normal univariada;

vs

$H_1 : \left(\hat{\beta}_k^{*(1)}, ..., \hat{\beta}_k^{*(B)}\right)$ não provêm de uma distribuição normal univariada.

Em cada um dos testes efectuados, começou por se considerar a dimensão da amostra de 100 réplicas *bootstrap*, ou seja, $B = 100$. Mesmo com este número de réplicas relativamente baixo, os resultados, que são apresentados na Tabela 5.1, foram bastante satisfatórios. Tendo em conta a dimensão da amostra, utilizou-se o teste de ajustamento de Lilliefors.

As conclusões foram tomadas tendo em conta a distribuição empírica dos valores-p devolvidos pelos testes. Note-se que a hipótese da distribuição normal não é rejeitada para valores-p superiores ao nível de significância do teste.

A Figura 5.3 mostra os diagramas de extremos e quartis das distribuições empíricas dos valores-p obtidos, para as grelhas de dimensão 10×10, 30×30 e 50×50. Os resultados estão apresentados, para cada estimador-MM $\hat{\beta}_0, \hat{\beta}_1$ e $\hat{\beta}_2$, em função do lado da grelha. Na figura é possível observar que a grande maioria dos valores-p obtidos, aponta para a não rejeição da distribuição normal. Como era de esperar, à medida que a dimensão da grelha aumenta, o número de valores-p pequenos vai diminuindo, isto é, a percentagem de rejeições da normalidade diminui.

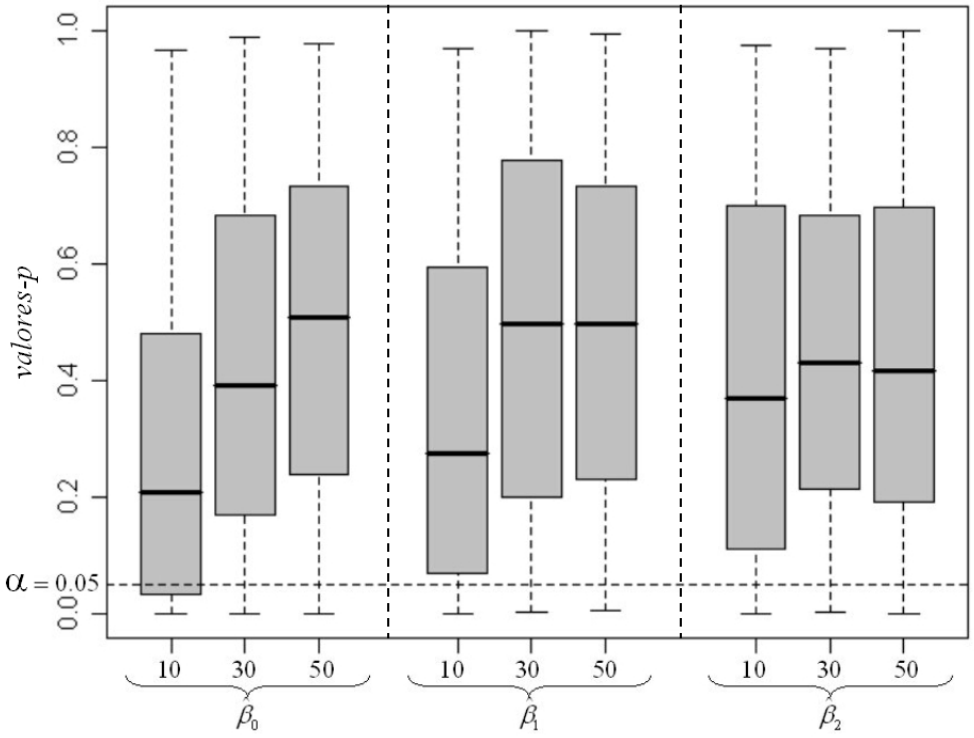

Figura 5.3: Diagramas de extremos e quartis das distribuições empíricas dos valores--p obtidos através dos testes de Lilliefors, para as grelhas de dimensão 10×10, 30×30 e 50×50.

Na Tabela 5.1 está representada a percentagem de testes de Lilliefors que não rejeitaram a hipótese nula, para um nível de significância $\alpha = 5\%$. Enquanto que na Figura 5.3 apenas se reproduzem os resultados relativos a três dimensões de grelha, na

Tabela 5.1 são apresentados todos os casos considerados.

	10×10	20×20	30×30	40×40	50×50	60×60
$\hat{\beta}_0$	69 %	74 %	87 %	97 %	92 %	96 %
$\hat{\beta}_1$	80 %	82 %	88 %	94 %	95 %	95 %
$\hat{\beta}_2$	83 %	85 %	95 %	93 %	91 %	96 %

Tabela 5.1: Percentagem de testes de Lilliefors que não rejeitaram a hipótese da distribuição normal das réplicas *bootstrap*, para um nível de significância de $\alpha = 5\%$.

Da análise da Tabela 5.1 observa-se que, à medida que a dimensão da grelha aumenta, o número de rejeições da distribuição normal vai diminuindo, tal como já se podia concluir pela Figura 5.3.

Chama-se à atenção para o facto de se demonstrar que, à medida que a dimensão da grelha aumenta, os valores da Tabela 5.1 deverem convergir para a probabilidade desses mesmos valores-p serem superiores ao nível de significância α. De facto, interpretando o valor-p de um teste como uma variável aleatória, sabe-se que ela segue uma distribuição uniforme no intervalo $]0, 1[$, supondo que se verifica a hipótese nula do referido teste. Assim, os valores da Tabela 5.1 devem convergir para a probabilidade de uma variável aleatória com distribuição uniforme em $]0, 1[$ ser superior a $\alpha = 5\%$, isto é, os valores devem convergir para $1 - \alpha = 95\%$. Efectivamente, os resultados da Tabela 5.1 parecem confirmar as expectativas, uma vez que se aproximam rapidamente dos 95%, em qualquer uma das linhas da tabela.

Outro pormenor que é de realçar é que, para cada estimador fixo, os diagramas de extremos e quartis dos valores-p (Figura 5.3) tornam-se cada vez mais simétricos à medida que a dimensão da grelha aumenta; os quartis também se aproximam cada vez mais do seu valor teórico na distribuição uniforme em $]0, 1[$.

Como conclusão, não existem motivos para rejeitar a hipótese nula, aceitando-se que a distribuição assintótica dos estimadores-MM, sob *IDA*, pode ser aproximada por uma distribuição normal.

$\Diamond$

5.3 Um teste à estacionaridade da média

A estacionaridade da média, como o próprio nome indica, é a propriedade que garante que a média de um processo $Z(\mathbf{s})$ é constante em todo o seu domínio $D \subset \mathbb{R}^d$, *i.e.*,

$$\forall_{\mathbf{s}\in D} \quad \mathrm{E}[Z(\mathbf{s})] = \mu \in \mathbb{R}.$$

Como se pode ver na secção **1.2**, sempre que as variáveis aleatórias do processo têm momentos de primeira ordem, qualquer um dos tipos de estacionaridade que são geralmente considerados em Geoestatística assumem que esta propriedade se verifica. No entanto, é fácil encontrar processos geoestatísticos onde a estacionaridade da média é questionável. Por exemplo, quando se observa a temperatura dos países da Europa, é claro que à medida que se avança para norte, o valor que se espera obter nas temperaturas registadas vai diminuindo. Por isso, é abusivo considerar que a média das temperaturas registadas é constante em toda a Europa.

O facto de se ignorar a falta de estacionaridade da média constitui um problema para a modelação geoestatística, pois se a média de um processo não for constante, todos os procedimentos usuais de estimação perdem as suas boas propriedades.

Para confirmar a afirmação anterior, considere-se um processo $Z(\mathbf{s}) = \mu(\mathbf{s}) + \delta(\mathbf{s})$, onde $\mu(\mathbf{s})$ é uma função determinística não constante (a chamada *tendência*), que varia com a localização $\mathbf{s} \in D \subset \mathbb{R}^d$, e $\delta(\mathbf{s})$ é um qualquer processo estacionário de segunda ordem com média nula. Suponha-se que a falta de estacionaridade da média presente no processo $Z(\mathbf{s})$ é ignorada e que se passa directamente à estimação pontual do variograma através do estimador de Matheron definido em (2.1.1). Deste modo, para cada vector $\mathbf{h} \in \mathbb{R}^d$ onde foi estimado o variograma, tem-se que

$$\begin{aligned}
\mathrm{E}[2\hat{\gamma}(\mathbf{h})] &= \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} \mathrm{E}\left[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2\right] \\
&= \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} \left\{\mathrm{Var}[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)] + \mathrm{E}[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]^2\right\} \\
&= 2\gamma(\mathbf{h}) + \frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2.
\end{aligned}$$

Isto quer dizer que o estimador usual do variograma deixou de ser centrado, tendo um enviesamento de $\frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} [\mu(\mathbf{s}_i) - \mu(\mathbf{s}_j)]^2$. Este enviesamento vai afectar a estimação do variograma e, consequentemente, todos os procedimentos subsequentes.

Sendo assim, torna-se imprescindível detectar as situações onde a estacionaridade da média não se verifica de modo significativo, para que se possam adoptar procedimentos adequados. Por exemplo, quando se detecta que um dado processo não tem a média constante, existem métodos conhecidos para remover a tendência, tornando a média constante em todo o domínio. Com isto obtém-se um processo estacionário, onde é possível utilizar os métodos de inferência tradicionais. No final da análise estatística, volta-se a adicionar a tendência, para se regressar ao processo original. Um dos principais métodos que tem por base esta ideia é designado por *median polish*. A sua descrição detalhada pode-se encontrar, por exemplo, em Cressie (1993).

No entanto, a decisão de considerar que existe (ou não) tendência é tomada com base em critérios subjectivos. Cressie (1993) propôs algumas técnicas de análise preliminar de dados, que podem ajudar a detectar a ausência da estacionaridade da média, em processos onde o domínio é um subconjunto do plano $\mathbb{R}^2$. Essa proposta aplica-se a observações dispostas ao longo de grelhas regulares, e consiste em analisar as médias e as medianas das linhas e/ou das colunas de observações do processo. Assim, caso se pretenda fazer a analise da média do processo na direcção das linhas, calcula-se a média e a mediana amostrais das observações de cada coluna. Depois averigua-se como estas variam de coluna para coluna, ou seja, na direcção das linhas. Isto permite detectar alguma tendência ao longo das linhas. Pode-se utilizar o mesmo raciocínio para averiguar se existe alguma tendência na direcção das colunas.

Segundo Cressie (1993), o facto das observações não estarem dispostas ao longo de uma grelha regular não é impeditivo de se fazer esta análise. Para tal, basta aproximar as localizações por pontos de uma grelha regular, considerando que cada localização pertence à linha e à coluna que lhe estão mais próximas.

Este tipo de análise permite ficar com uma ideia de como varia a média de um processo de domínio bidimensional na direcção das suas linhas e/ou colunas. No entanto, a decisão continua a ser subjectiva, uma vez que depende da opinião de quem está a fazer a análise. Por outro lado, o principal objectivo de Cressie quando sugere este método, é o de detectar a presença de observações atípicas na amostra, tendo em conta a eventual diferença entre a média e a mediana amostrais. A análise não é aproveitada para efectuar um teste estatístico à estacionaridade da média.

Tendo em conta a importância da existência de um teste que permita confirmar a estacionaridade da média, de seguida desenvolve-se uma proposta que tem por base um método de projecções. Esse método é semelhante ao atrás referido como sugestão de Cressie para detectar observações atípicas; vai ser útil para reformular a questão da estacionaridade da média.

5.3.1 A metodologia do teste

Considere-se um processo univariado $Z(\mathbf{s})$ com domínio $D \subset \mathbb{R}^d$. Para facilitar a apresentação, vai-se considerar que o domínio D é um subconjunto do plano $\mathbb{R}^2$. No entanto, este teste pode ser generalizado a outras dimensões, nomeadamente a $\mathbb{R}^3$.

Ao longo do teste será utilizada a noção de m-dependência, a qual é expressa em termos da covariância. Por esse motivo, suponha-se ainda que a estrutura de dependência de $Z(\mathbf{s})$ é traduzida por um covariograma isotrópico

$$C(\mathbf{h}) = \mathrm{Cov}[Z(\mathbf{s}+\mathbf{h}), Z(\mathbf{s})],\ \forall_{\mathbf{s}\in D},\ \forall_{\mathbf{h}\in\mathbb{R}^2} \text{ tal que } \mathbf{s}+\mathbf{h} \in D.$$

Como foi referido na subsecção **1.3.1**, uma vez que $Z(\mathbf{s})$ tem covariograma, então ele apresenta, no mínimo, uma amplitude prática $m \in [0, +\infty[$; por outro lado, ao assumir que o covariograma é isotrópico, m é constante em todas as direcções. Portanto, se duas variáveis aleatórias do processo $Z(\mathbf{s})$ estão separadas por uma distância superior a m, elas são consideradas não correlacionadas, isto é, o processo $Z(\mathbf{s})$ é considerado m-dependente (*vide* **Definição 5.1.1**).

Dado um processo $Z(\mathbf{s})$ que verifica as hipóteses referidas anteriormente, apresenta-se um teste para confirmar se a média de $Z(\mathbf{s})$ é constante em todo o domínio D (ou seja, se a estacionaridade da média se verifica), contra a hipótese alternativa de que existe uma tendência polinomial em $Z(\mathbf{s})$. Note-se que, como qualquer função contínua pode ser aproximada por uma função polinomial, apenas se consideram tendências do tipo polinomial na hipótese alternativa. Formalmente, o que se pretende é efectuar o seguinte teste de hipóteses

$H_0 : \forall_{\mathbf{s}_i,\mathbf{s}_j\in D}\ \mathrm{E}[Z(\mathbf{s}_i)] = \mathrm{E}[Z(\mathbf{s}_j)] = \mu \in \mathbb{R}$;

vs

$H_1 : \forall_{\mathbf{s}\in D}\ \mathrm{E}[Z(\mathbf{s})] = g(\mathbf{s})$, onde g é uma função polinomial não constante.

Para efectuar o teste considera-se uma amostra $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n)\}$ do processo $Z(\mathbf{s})$, cujas localizações estão dispostas ao longo de uma grelha regular do plano, tal como mostra a Figura 5.1. O teste é efectuado com o auxílio de um método de projecções, o qual transforma as observações $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n)\}$ da amostra em pares ordenados $\{(x_1, Z_1), ..., (x_n, Z_n)\}$. Ao tomar projecções, o teste é enquadrado num problema de regressão, o que permite utilizar estimadores robustos. Em particular, permite a utilização do LAD e dos estimadores-MM, os quais foram estudados nas secções **5.1** e **5.2**.

5.3.2 O método das projecções

Na presente subsecção sugere-se a utilização de um procedimento baseado em projecções ortogonais, de modo a transformar as localizações das observações da amostra, que são multidimensionais, em números reais. A ideia fundamental deste método reside no facto de que, se existir uma tendência polinomial em $Z(\mathbf{s})$, então existe pelo menos uma recta l onde, para qualquer localização $\mathbf{s} \in l$, a função

$$\begin{array}{rccl} g\colon & l \subset \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & \mathbf{s} & \longmapsto & g(\mathbf{s}) = \mathrm{E}[Z(\mathbf{s})] \end{array}$$

é uma função polinomial não constante. Assim, para averiguar se existe uma tendência polinomial num dado processo $Z(\mathbf{s})$ (a partir de uma amostra $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), ..., Z(\mathbf{s}_n)\}$), poder-se-ia testar, para todas as rectas definidas pelas localizações $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ da amostra, se a função g é, ou não, uma função polinomial não constante. No entanto, este procedimento não seria adequado, uma vez que o número de rectas definidas pelas localizações da amostra pode ser bastante grande e, para além disso, cada uma dessas rectas, geralmente, contém um número muito pequeno de localizações/observações, o que torna praticamente impossível a realização de qualquer teste.

Uma maneira de ultrapassar esta dificuldade consiste em agrupar todas as rectas que têm a mesma direcção, numa única recta, à custa de projecções ortogonais. Este procedimento evita uma análise separada de cada uma das rectas e permite juntar um número elevado de observações por cada direcção, por forma a que seja possível efectuar um teste em cada uma das direcções consideradas. A ideia é semelhante à de Cressie (1993) que foi apresentada na secção **5.3**. No entanto, Cressie apenas considerou

projecções ortogonais na direcção das linhas/colunas da grelha das localizações. Na metodologia que aqui se propõe, considera-se um conjunto de direcções com interesse e faz-se uma projecção ortogonal em cada uma dessas direcções. As projecções ortogonais vão permitir que se faça o teste à estacionaridade da média, em cada uma das direcções pretendidas.

Para facilitar a apresentação do método das projecções, é necessário introduzir alguma notação. Assim, considere-se a direcção do vector unitário $\vec{\mathbf{e}} = \mathbf{h}_0 / \|\mathbf{h}_0\|$, e denote-se por $l_{\vec{\mathbf{e}}}$ uma recta representativa de todas as rectas com a direcção de $\mathbf{h}_0$. Sejam $x_1, ..., x_n$ os pontos obtidos através da projecção ortogonal das localizações $\mathbf{s}_1, ..., \mathbf{s}_n$ sobre a recta $l_{\vec{\mathbf{e}}}$ e denotem-se por Z_i as variáveis aleatórias $Z(\mathbf{s}_i)$, para $i = 1, ..., n$.

A Figura 5.4 ilustra o método das projecções ao longo da direcção do vector $\vec{\mathbf{e}}$. Do lado esquerdo da figura, é possível visualizar a localização $\mathbf{s}_i$ e a sua correspondente projecção ortogonal x_i sobre a recta $l_{\vec{\mathbf{e}}}$. Do lado direito da mesma figura, é possível ver como as projecções ortogonais das observações, dão origem a uma nova amostra, na qual se pode definir o modelo de regressão das observações Z_i sobre as projecções x_i das localizações ($i = 1, ..., n$). A representação gráfica das projecções num diagrama de dispersão também tem a vantagem de, numa análise preliminar, sugerir formas específicas para $\mathrm{E}[Z(\mathbf{s})]$.

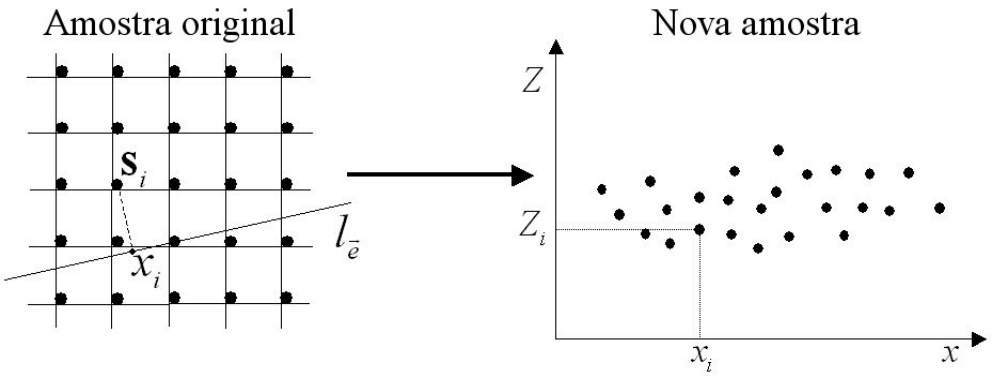

Figura 5.4: Ilustração do método das projecções. Com as projecções ortogonais das localizações da amostra original sobre a recta $l_{\vec{\mathbf{e}}}$, forma-se uma nova amostra onde se pode definir um modelo de regressão.

O conjunto dos pares ordenados $\{(x_1, Z_1), ..., (x_n, Z_n)\}$ da nova amostra permite testar se a função $g_{\vec{\mathbf{e}}}\colon \mathbb{R} \to \mathbb{R}$, que é definida por $g_{\vec{\mathbf{e}}}(x) = \mathrm{E}[Z(x)|x]$, é ou não uma função polinomial constante. Para cada direcção fixa do vector $\vec{\mathbf{e}}$, é possível efectuar esse teste recorrendo a um modelo de regressão linear, onde

$$g_{\vec{\mathbf{e}}}(x) = \mathrm{E}[Z(x)|x] = \beta_0 + \sum_{k=1}^{p} \beta_k x^k,\ p \in \mathbb{N}, \tag{5.3.1}$$

com $Z_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_i^k + \epsilon_i$ e $\mathrm{E}[\epsilon_i|x_i] = 0$, para $i = 1, ..., n$.

O teste é efectuado testando o valor dos coeficientes do modelo de regressão, isto é,

$$H_0 : \forall_{k\in\{1,...,p\}} \quad \beta_k = 0 \qquad vs \qquad H_1 : \exists_{k\in\{1,...,p\}} \quad \beta_k \neq 0. \tag{5.3.2}$$

Se, em alguma das direcções consideradas, existirem motivos para rejeitar a hipótese nula, então é porque a função $g_{\vec{\mathbf{e}}}(x)$ não é constante ao longo dessa direcção. Isto significa que existem pelo menos dois pontos $\mathbf{s}_i$ e $\mathbf{s}_j$ pertencentes ao domínio D, tais que $E[Z(\mathbf{s}_i)] \neq E[Z(\mathbf{s}_j)]$, o que leva a concluir que o processo $Z(\mathbf{s})$ não apresenta estacionaridade da média. Se, por outro lado, a hipótese nula não for rejeitada em qualquer uma das direcções consideradas, poder-se-á aceitar que o processo $Z(\mathbf{s})$ apresenta estacionaridade da média. É de salientar que não interessa testar o parâmetro β_0, uma vez que este não contribui para a existência (ou não) de tendência.

Como se supõe que as localizações da amostra do processo $Z(\mathbf{s})$ são fixas, então as suas projecções ortogonais x_i também o são. Como o processo $Z(\mathbf{s})$ é considerado m-dependente, então as observações Z_i e, consequentemente, os erros ϵ_i do modelo de regressão, também o vão ser $(i = 1, ..., n)$. Portanto, o método das projecções transforma o teste da estacionaridade da média, num conjunto de testes efectuados em modelos de regressão linear, com regressores fixos e erros m-dependentes. Sendo assim, o modelo de regressão linear caracterizado por (5.3.1) é idêntico ao do ponto (5.1.1), que foi considerado nas secções **5.1** e **5.2**. Por isso, os estimadores apresentados nessas duas secções, podem ser utilizados para efectuar o teste (5.3.2).

Contudo, tal como se referiu anteriormente, as matrizes de covariâncias, quer dos estimadores-MM, quer do estimador *LAD*, são difíceis de estimar através da metodologia tradicional, dada a estrutura de dependência revelada pelos erros do modelo. Deste modo, seguidamente apresenta-se uma metodologia *bootstrap*, baseada no método das

projecções, que permite aproximar a matriz de covariâncias de $\hat{\boldsymbol{\beta}}_{LAD}$ e de $\hat{\boldsymbol{\beta}}_{MM}$. Como a metodologia é a mesma para ambos os estimadores, a partir de agora, $\hat{\boldsymbol{\beta}}$ vai representar, quer $\hat{\boldsymbol{\beta}}_{LAD}$, quer $\hat{\boldsymbol{\beta}}_{MM}$.

A metodologia *bootstrap* que se utiliza para aproximar a matriz de covariâncias de $\hat{\boldsymbol{\beta}}$ é a *SFBB* apresentada na secção **5.2** para aproximar a distribuição assintótica dos estimadores-MM, sob condições de m-dependência. No entanto, para recorrer às projecções, existe uma etapa intermédia na sua aplicação (em relação ao que foi descrito anteriormente). Logo, todos os argumentos que são válidos para justificar a utilização da metodologia *bootstrap SFBB* da secção **5.2** vão permanecer válidos para esta secção.

Assim, os blocos *bootstrap* continuam a ser formados com as observações do processo original $Z(\mathbf{s})$, uma vez que são essas observações que revelam a estrutura de dependência original. Depois de formar um bloco de observações do processo original através da distância d' definida em (5.2.1), tal como foi explicado na secção **5.2**, utiliza-se o método das projecções para transformar esse bloco, num outro bloco formado por observações do modelo de regressão linear, tal como mostra a Figura 5.5.

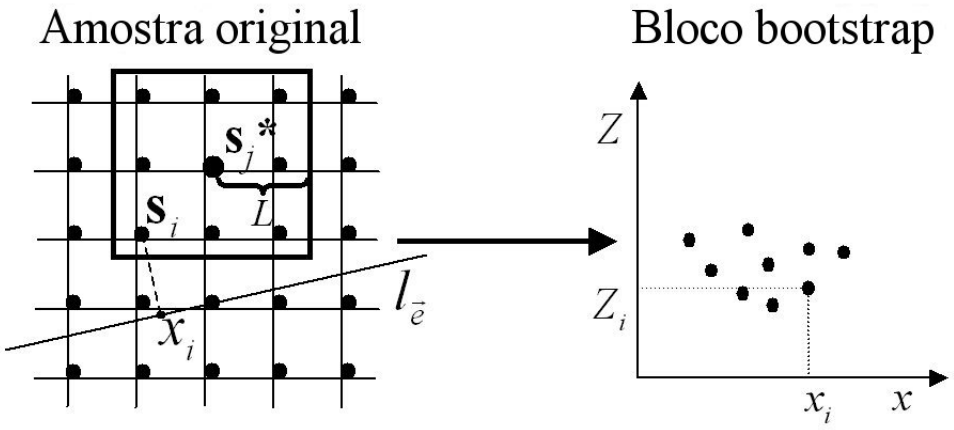

Figura 5.5: Esquema da construção dos blocos *bootstrap*.

Resumindo, depois de seleccionar aleatoriamente a observação $\mathbf{s}_j^*$ e de formar o bloco sobre a amostra original do processo $Z(\mathbf{s})$, utiliza-se o método das projecções para obter o bloco associado, o qual é formado pelas observações (x_i, Z_i), onde x_i é a projecção ortogonal de $\mathbf{s}_i$ sobre a recta $l_{\vec{\mathbf{e}}}$ e Z_i é igual à correspondente variável aleatória

$Z(\mathbf{s}_i)$ da amostra de $Z(\mathbf{s})$. Cada bloco é então formado por

$$b_j^* = \{(x_i, Z(\mathbf{s}_i)) : d'(\mathbf{s}_j^\star, \mathbf{s}_i) \leq L\}. \tag{5.3.3}$$

A metodologia *bootstrap* que se utiliza para aproximar a matriz de covariâncias $\mathbf{V}_n$ do estimador $\hat{\boldsymbol{\beta}}$, pode ser sumariada nos seguintes passos:

1. Escolhe-se, ao acaso, uma das localizações $\{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ da grelha da amostra original de $Z(\mathbf{s})$ e denomina-se essa localização de $\mathbf{s}_j^*$.

2. Com a localização $\mathbf{s}_j^*$ forma-se o bloco $b_j^* = \{(x_i, Z(\mathbf{s}_i)) : d'(\mathbf{s}_j^*, \mathbf{s}_i) \leq L, 0 < L \leq m\}$, onde $d'(\mathbf{s}_i, \mathbf{s}_j) = \max_{k=1,2} |s_{i,k} - s_{j,k}|$.

3. Para fixar o número de observações por bloco, consideram-se apenas blocos com $n_o = \max_{j=1,...,n_b}\{\#b_j^*\}$ observações. Caso a observação seleccionada dê origem a um bloco com menos observações do que n_o, ignora-se essa observação e selecciona-se outra.

4. Consideram-se blocos b_j^* sucessivos até se obter a amostra *bootstrap* $(b_1^*, ..., b_{n_b}^*)$. O número total de blocos n_b é o menor inteiro que satisfaz a condição $n_b \times n_o \geq n$.

5. A partir da amostra *bootstrap* obtém-se uma réplica *bootstrap* $\hat{\beta}_k^{*(b)}$ do estimador $\hat{\beta}_k$, para $k = 1, ..., p$.

6. Repetem-se os passos anteriores B vezes obtendo-se um conjunto de réplicas *bootstrap* $\left(\hat{\beta}_k^{*(1)}, ..., \hat{\beta}_k^{*(B)}\right)$ do estimador de β_k, para $k = 1, ..., p$.

Como se referiu, em relação à secção **5.2**, apenas foi alterado o passo número 2, por forma a incorporar o método das projecções.

Com as réplicas *bootstrap* $\left(\hat{\beta}_k^{*(1)}, ..., \hat{\beta}_k^{*(B)}\right)$, é possível obter uma estimativa da matriz de covariâncias do estimador $\hat{\boldsymbol{\beta}}, \mathbf{V}_n$. Deste modo, para qualquer $k = 1, ..., p$ e $l = 1, ..., p$, o elemento da k-ésima linha e da l-ésima coluna de $\mathbf{V}_n$ é dado por

$$\widehat{\mathbf{V}}_n[k, l] = \widehat{\mathrm{Cov}}[\hat{\beta}_k, \hat{\beta}_l] = \frac{1}{B}\sum_{j=1}^{B}(\hat{\beta}_k^{\star(j)} - \hat{\beta}_k)(\hat{\beta}_l^{\star(j)} - \hat{\beta}_l). \tag{5.3.4}$$

No caso particular de k ser igual a l, isto é, na diagonal principal da matriz $\widehat{\mathbf{V}}_n$, encontram-se as estimativas da variância dos estimadores $\hat{\beta}_k$, as quais vão ser denotadas por $\hat{s}^2_{\hat{\beta}_k}$.

Deste modo, ficam reunidas as condições para que seja possível aplicar os resultados estabelecidos nas secções **5.1** e **5.2**, sobre a distribuição assintótica do LAD e dos estimadores-MM. Então, o teste (5.3.2) pode ser efectuado recorrendo ao facto de

$$\hat{\boldsymbol{\beta}} \xrightarrow{\mathcal{L}} N_p(\boldsymbol{\beta}, \widehat{\mathbf{V}}_n). \tag{5.3.5}$$

Note-se que, deste modo, a questão de testar a estacionaridade da média é recolocada no contexto de testes paramétricos em populações normais, o que facilita a sua aplicação/divulgação.

5.3.3 Algoritmo do teste

O teste à estacionaridade da média efectua-se quando, numa análise preliminar de dados, se suspeitar da existência de tendência no processo $Z(\mathbf{s})$. Quando isto acontece, deve-se encontrar a direcção, ou as direcções, onde essa tendência é mais pronunciada e efectuar o teste (5.3.2) ao longo dessas direcções. A região crítica do teste vai depender, quer do grau p ($p \geq 1$) do polinómio colocado na hipótese alternativa, quer do número de direcções $N \in \mathbb{N}$ onde se pretende efectuar o teste. Sendo assim, nesta secção apresentam-se, sucessivamente, as regiões críticas do teste à estacionaridade da média, para diferentes valores de N e de p.

A primeira situação estudada é a mais simples e corresponde ao caso em que há a suspeita de uma tendência linear ($p = 1$) numa única direcção – pretende-se efectuar o teste (5.3.2) apenas numa direcção, ou seja, para $p = N = 1$.

Neste caso, faz-se um teste do tipo

$$H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0,$$

na direcção escolhida.

Assumindo a distribuição assintótica normal do estimador $\hat{\beta}_1$, a região crítica do teste, a um nível de significância α ($0 < \alpha < 1$), é formada pelas amostras que dão origem a grandes valores absolutos de $\hat{\beta}_1$, mais precisamente,

$$RC_{1,1} = \left\{ (z(\mathbf{s}_1), ..., z(\mathbf{s}_n)) : \left| \frac{\hat{\beta}_1}{\hat{s}_{\hat{\beta}_1}} \right| > q_{1-\alpha/2} \right\}, \tag{5.3.6}$$

onde $\hat{s}_{\hat{\beta}_1}$ é a estimativa *bootstrap* do desvio padrão de $\hat{\beta}_1$ determinada por (5.3.4) e $q_{1-\alpha/2}$ é o quantil $1-\alpha/2$ da distribuição normal standardizada.

Considere-se agora que se pretende averiguar se existe uma tendência linear ($p=1$) em mais do que uma direcção, isto é, que se pretende efectuar o teste (5.3.2) para $p=1$ e $N>1$ finito.

Nesta situação, é necessário verificar se existe tendência linear em todas as N direcções consideradas. Por isso é preciso testar

$$H_{0,i}: \beta_{1,i}=0 \quad vs \quad H_{1,i}: \beta_{1,i}\neq 0,$$

em cada direcção $i=1,...,N$. Tendo em conta que há interesse em testar simultaneamente as N hipóteses garantindo o nível de significância α, há a necessidade de utilizar o método de reunião-intersecção, o qual se passa a descrever.

No teste que se está a considerar, apenas se aceita que o processo $Z(\mathbf{s})$ tem estacionaridade da média quando todas as hipóteses $H_{0,i}$ não são rejeitadas ou, por outras palavras, só se rejeita a hipótese nula quando pelo menos uma das $H_{0,i}$ é rejeitada. Portanto, a hipótese nula do teste para todas as direcções é a conjunção das hipóteses nulas de cada uma das direcções consideradas. Assim, o teste é explicitado como

$$H_0: \bigwedge_{i=1}^{N}(\beta_{1,i}=0) \quad vs \quad H_1: \bigvee_{i=1}^{N}(\beta_{1,i}\neq 0).$$

Logo, a região crítica do teste é formada por todas as observações que fazem com que pelo menos um dos $\hat{\beta}_{1,i}$ tenha um valor absoluto elevado, ou seja, por

$$RC_{1,N}=\left\{(z(\mathbf{s}_1),...,z(\mathbf{s}_n)): \left|\frac{\hat{\beta}_{1,1}}{\hat{s}_{\hat{\beta}_{1,1}}}\right|>q_{1,N}\vee ...\vee \left|\frac{\hat{\beta}_{1,N}}{\hat{s}_{\hat{\beta}_{1,N}}}\right|>q_{1,N}\right\}. \tag{5.3.7}$$

Assumindo que as N direcções são escolhidas *a priori* de uma forma independente, então as estatísticas $\frac{\hat{\beta}_{1,1}}{\hat{s}_{\hat{\beta}_{1,1}}},...,\frac{\hat{\beta}_{1,N}}{\hat{s}_{\hat{\beta}_{1,N}}}$ também podem ser consideradas independentes. Consequentemente, os quantis $q_{1,N}$ são iguais para todas as N direcções e a região crítica $RC_{1,N}$ pode ser reescrita na forma

$$RC_{1,N}=\left\{(z(\mathbf{s}_1),...,z(\mathbf{s}_n)): \max_{i=1,...,N}\left|\frac{\hat{\beta}_{1,i}}{\hat{s}_{\hat{\beta}_{1,i}}}\right|>q_{1,N}\right\}.$$

Por outro lado, quando se verifica a hipótese nula, cada $\frac{\hat{\beta}_{1,i}}{\hat{s}_{\hat{\beta}_{1,i}}}, i = 1, ..., N$, segue uma distribuição assintótica normal standardizada. Então, o valor de $q_{1,N}$ é obtido através do quantil $1-\alpha$ da variável aleatória $\max\limits_{i=1,...,N} \left| \frac{\hat{\beta}_{1,i}}{\hat{s}_{\hat{\beta}_{1,i}}} \right|$, a qual tem $f.d.p.$ dada por

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ (2\Phi(x) - 1)^N & \text{if } x \geq 0 \end{cases},$$

onde, como é usual, $\Phi(x)$ representa a $f.d.p.$ normal standardizada.

Por fim, resta analisar o caso geral, correspondente à suspeita de uma tendência polinomial de grau $p > 1$ em N direcções.

Neste caso, o teste é explicitado através de

$$H_0 : \bigwedge_{k=1}^{p} \bigwedge_{i=1}^{N} (\beta_{k,i} = 0) \quad vs \quad H_1 : \bigvee_{k=1}^{p} \bigvee_{i=1}^{N} (\beta_{k,i} \neq 0).$$

Para determinar a região crítica do teste note-se que, para um i fixo ($i = 1, ..., N$), as distribuições de $\hat{\beta}_{1,i}, \hat{\beta}_{2,i}, ..., \hat{\beta}_{p,i}$ são geralmente dependentes, dado que são os estimadores dos coeficientes do modelo de regressão. Para além disso, a matriz de covariâncias de $\hat{\boldsymbol{\beta}}_i = (\hat{\beta}_{1,i}, \hat{\beta}_{2,i}, ..., \hat{\beta}_{p,i})$, geralmente é diferente de direcção para direcção.

Para ultrapassar estes problemas, considere-se o vector pN-dimensional $\hat{\boldsymbol{\beta}}_{pN}$ que é constituído por todos os vectores p-dimensionais $\hat{\boldsymbol{\beta}}_i$ de todas as N direcções, ou seja,

$$\hat{\boldsymbol{\beta}}_{pN} = vec(\hat{\boldsymbol{\beta}}_1, ..., \hat{\boldsymbol{\beta}}_N).$$

Sob a hipótese nula, o vector $\hat{\boldsymbol{\beta}}_{pN}$ tem uma distribuição assintótica normal multivariada, com vector médio nulo e matriz de covariâncias $\widehat{\mathbf{V}}$. A matriz de covariâncias $\widehat{\mathbf{V}}$ é uma matriz diagonal por blocos. Os blocos que se encontram na diagonal principal são as matrizes $\widehat{\mathbf{V}}_{n,i}$ denotadas por $\widehat{\mathbf{V}}_n$ em (5.3.4).

Os quantis da região crítica do teste são determinados tendo em conta que

$$\widehat{\mathbf{V}}^{-\frac{1}{2}} \hat{\boldsymbol{\beta}}_{pN} \xrightarrow{\mathcal{L}} N_{pN}(\mathbf{0}, \mathbf{I}_{pN}),$$

onde $\mathbf{I}_{pN}$ representa a matriz identidade de ordem pN.

Neste caso, também se pode efectuar um teste para cada $\beta_{k,i}$, separadamente ($k = 1, ..., p$ e $i = 1, ..., N$). Quando se opta por essa via, é necessário fazer uma correcção ao nível de significância de cada teste, de acordo com o nível de significância pretendido para o teste global. Para tal, pode-se utilizar, por exemplo, a correcção de Bonferroni.

5.3.4 Exemplo de aplicação a um conjunto de dados reais

Nesta secção ilustra-se a aplicação do teste à estacionaridade da média, proposto na secção anterior, num contexto de observações reais. Os dados foram recolhidos num montado de Portel (Alto Alentejo, Portugal), durante o projecto europeu MEDAFOR (ENV4-CT97-0686), cujos resultados estão publicados em Shakesby, Coelho, Schnabel, Keizer, Clarke, Contador, Walsh, Ferreira e Doerr (2002). Um dos objectivos do projecto era o estudo da humidade volumétrica do solo. Sendo assim, o processo $Z(\mathbf{s})$ que a seguir se analisa, representa a percentagem de humidade do solo presente na localização $\mathbf{s} \in D \subset \mathbb{R}^2$. Na Figura 5.6 pode-se visualizar a amostra recolhida, a qual é constituída por 180 observações, dispostas ao longo de uma grelha regular do plano, de 20 linhas e 9 colunas. Quer as linhas, quer as colunas da grelha, encontram-se separadas por uma distância de 5 metros. Isto significa que as localizações estão dispostas nos vértices de quadrados com 5 metros de lado.

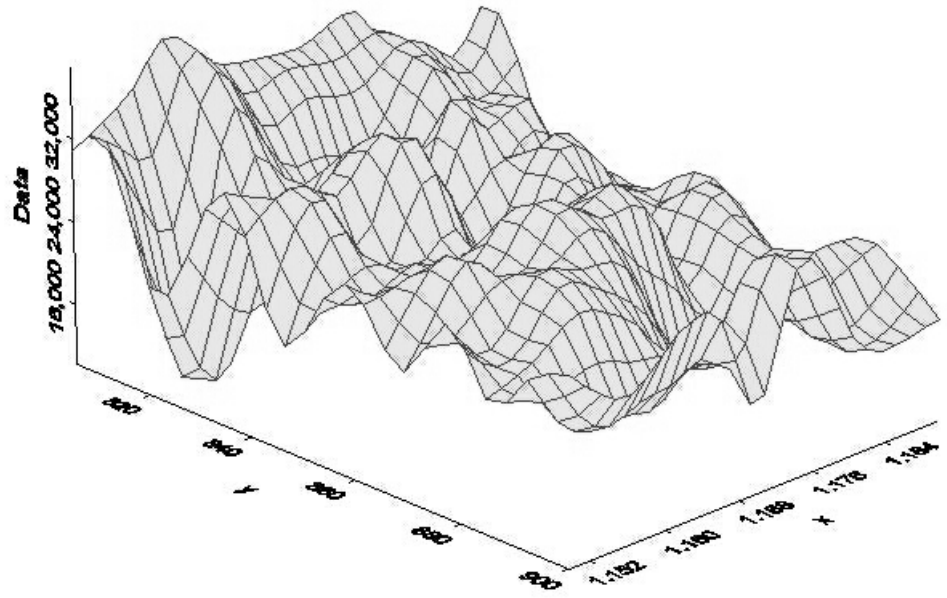

Figura 5.6: Representação da amostra do processo de concentrações de humidade do solo.

A Figura 5.6 mostra que, na direcção do vector $\vec{i} = (1, 0)$ (direcção do eixo Ox), parece não haver motivos para colocar em causa a estacionaridade da média. Contudo, isto não é tão óbvio na direcção de $\vec{j} = (0, 1)$ (direcção do eixo Oy). Esta última

direcção, leva mesmo a pensar que existe uma tendência linear em $Z(\mathbf{s})$. Consequentemente, decidiu-se efectuar um teste à estacionaridade da média do processo $Z(\mathbf{s})$, tendo em conta as duas direcções principais de $\vec{i}$ e de $\vec{j}$ que são definidas pela grelha das localizações. Na hipótese alternativa supõe-se que existe uma tendência do tipo linear, traduzida por uma recta de regressão, logo $p = 1$. Portanto, a região crítica do teste será da forma (5.3.7), com $N = 2$.

Para efectuar o teste, estimou-se o coeficiente de regressão β_1, ou seja, o declive, em ambas as direcções, quer usando o LAD, quer usando um estimador-MM.

Na direcção de $\vec{i}$ obteve-se $\hat{\beta}_{LAD} = -0.076$ e $\hat{\beta}_{MM} = -0.072$. Na direcção de $\vec{j}$ obteve-se $\hat{\beta}_{LAD} = -0.098$ e $\hat{\beta}_{MM} = -0.081$.

As estimativas foram obtidas através do software *R*. Utilizou-se a *package quantreg* para determinar $\hat{\beta}_{LAD}$ e a *package roblm* para determinar $\hat{\beta}_{MM}$. Ambas as *packages* foram utilizadas com as opções existentes por defeito.

Na Figura 5.7 podem-se visualizar as projecções ortogonais em ambas as direcções consideradas e as correspondentes rectas de regressão, obtidas através do estimador *LAD* (linha a tracejado) e do estimador-MM (linha a cheio). As estimativas *LAD* e

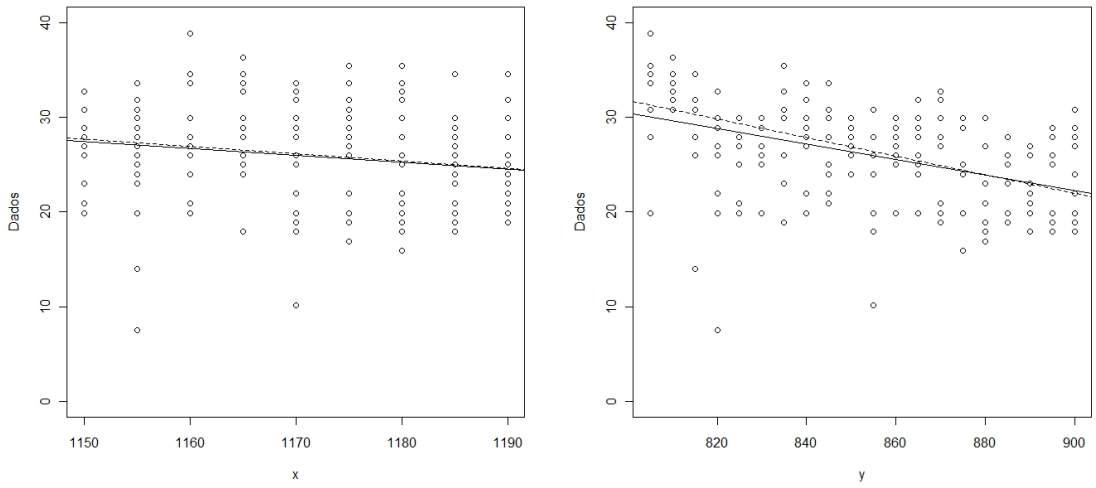

Figura 5.7: Diagramas de dispersão das projecções na direcção do vector $\vec{i} = (1, 0)$ (à esquerda) e na direcção do vector $\vec{j} = (0, 1)$ (à direita). As rectas a tracejado foram obtidas com o estimador *LAD* e as rectas a cheio com o estimador-MM.

as estimativas MM encontram-se bastante próximas, e ambas parecem indicar que não

existe estacionaridade da média em $Z(\mathbf{s})$. Resta agora determinar os desvios padrão dos estimadores, para se poder tirar a conclusão do teste.

Para estimar $\hat{s}_{\hat{\beta}_{1,\vec{i}}}$ e $\hat{s}_{\hat{\beta}_{1,\vec{j}}}$ é preciso utilizar a metodologia *bootstrap SFBB*. Como no projecto MEDAFOR a amplitude de $Z(\mathbf{s})$ foi estimada em 5.3 metros, considerou-se que $Z(\mathbf{s})$ é um processo m-dependente, com $m = 5.3$. Deste modo, para preservar a estrutura de dependência dentro de cada um dos blocos *bootstrap* em (5.3.3), considerou-se que os blocos seriam formados com $L = m = 5.3$. Com o estimador *bootstrap* obteve-se, para o estimador LAD, $\hat{s}_{\hat{\beta}_{1,\vec{i}}} = 0.067$ e $\hat{s}_{\hat{\beta}_{1,\vec{j}}} = 0.039$, e para o estimador-MM obteve-se $\hat{s}_{\hat{\beta}_{1,\vec{i}}} = 0.059$ e $\hat{s}_{\hat{\beta}_{1,\vec{j}}} = 0.023$.

O ponto crítico da região crítica (5.3.7) para o nível de significância $\alpha = 5\%$ é aproximadamente igual a $q_{1,2} = 2.24$. Consequentemente verifica-se que, na direcção de $\vec{j}$, e para o estimador LAD,

$$\left| \frac{\hat{\beta}_{LAD}}{\hat{s}_{\hat{\beta}_{1,\vec{j}}}} \right| = \left| \frac{-0.098}{0.039} \right| \approx 2.5 > 2.24.$$

Por outro lado, na direcção de $\vec{j}$, e para o estimador-MM tem-se que

$$\left| \frac{\hat{\beta}_{MM}}{\hat{s}_{\hat{\beta}_{1,\vec{j}}}} \right| = \left| \frac{-0.081}{0.023} \right| \approx 3.5 > 2.24.$$

Apresentando ainda o mesmo resultado em termos do valor-p, com o estimador LAD obteve-se um valor-p de 2.5%, e com o estimador-MM obteve-se um valor-p de aproximadamente 0%.

Portanto, quer com o estimador LAD, quer com o estimador-MM conclui-se que, com base na amostra recolhida e com o nível de significância 5%, encontram-se motivos para rejeitar a hipótese nula, ou seja, fazendo o teste com qualquer um dos estimadores, é de rejeitar a hipótese de estacionaridade da média de $Z(\mathbf{s})$, aceitando-se que o processo $Z(\mathbf{s})$ possui uma tendência. Recomenda-se assim a utilização de um método de remoção da tendência, tal como a *median polish* sugerida em Cressie (1993).

Ainda recorrendo ao mesmo conjunto de observações, também é possível efectuar o teste à estacionaridade da média, utilizando apenas a direcção suspeita de $\vec{j}$. Neste caso, o procedimento é semelhante, mas a região crítica do teste vai ser da forma (5.3.6). Assim, com o estimador LAD obtém-se um valor-p do teste igual a 1.2%. Por outro lado, com o estimador-MM obtém-se um valor-p de aproximadamente 0%. Deste

modo, para o nível de significância de 5%, as conclusões do teste anterior mantêm-se para ambos os estimadores, isto é, quer com o estimador LAD, quer com o estimador-MM, com base na amostra recolhida, rejeita-se a existência de estacionaridade da média em $Z(\mathbf{s})$.

Capítulo 6

Estimação robusta do variograma

Neste capítulo faz-se um estudo detalhado da estimação do variograma, tendo em vista a obtenção de um estimador que consiga ser robusto e que, simultaneamente, consiga ter boa eficiência em modelos normais.

Em primeiro lugar, começa por se fazer notar a falta de robustez do procedimento tradicional na estimação do variograma. De seguida, apresentam-se algumas propostas robustas de estimação do variograma, já publicadas na literatura sobre o assunto. Contudo, essas propostas podem ser melhoradas em termos de eficiência, como se verá ao longo do presente capítulo.

Sendo assim, desenvolve-se uma metodologia baseada em múltiplos variogramas obtidos com a mesma amostra, que permite aumentar a eficiência dos estimadores robustos do variograma. O estimador proposto, que se vai designar por estimador de múltiplos variogramas, é estudado em relação às suas propriedades. O trabalho efectuado permitiu detectar situações, onde a identificabilidade dos parâmetros do variograma não está assegurada. Assim, estabelecem-se condições que garantem a unicidade de solução do estimador de múltiplos variogramas. Essas condições permanecem válidas, quando se estima o variograma pelo processo tradicional.

6.1 Estimação robusta e etapas de estimação do variograma

A grande maioria dos procedimentos usuais em inferência estatística tem boas propriedades no que diz respeito à consistência e à eficiência em modelos normais, mas

o mesmo não se verifica em relação à robustez. Em Geoestatística, esta constatação prevalece. De facto, o procedimento usual de estimação do variograma tem boas propriedades (*vide* **Capítulo 2**) mas não é um procedimento robusto.

Analisando a metodologia usual desse ponto de vista, é necessário considerar, separadamente, cada uma das etapas da estimação, concretamente:

- a obtenção das estimativas pontuais do variograma, através do estimador de Matheron, na primeira etapa;
- a estimação dos parâmetros do modelo de variograma, por mínimos quadrados, na segunda etapa.

Considere-se a primeira etapa da estimação. Como se pode ver no ponto (2.1.1), o estimador de Matheron é construído a partir de médias dos quadrados dos incrementos do processo. O facto das médias amostrais não serem robustas faz com que o estimador de Matheron também não seja robusto. Por isso, tal como a média amostral, o estimador pontual do variograma que é mais utilizado, tem ponto de ruptura nulo e função de influência ilimitada (como é realçado em Schabenberger e Gotway (2005)).

Como o ponto de ruptura do estimador de Matheron é nulo, basta existir uma observação atípica na amostra do processo $Z(\mathbf{s})$, para que as estimativas do variograma sejam afectadas. Porém, as consequências da estimação não robusta do variograma podem ser ainda mais graves do que quando se estima de forma não robusta outro tipo de parâmetros. Como exemplo, refira-se que o efeito de um único erro grosseiro na amostra é ampliado no cálculo dos incrementos, uma vez que cada observação contribui para o cálculo de diversos incrementos do processo. De facto, uma observação atípica numa amostra de n elementos entra no cálculo de $n-1$ incrementos; então essa observação atípica vai dar origem a $n-1$ incrementos atípicos. Deste modo, se o processo considerado tiver variograma isotrópico, uma observação atípica numa amostra de n observações, dá origem a $n-1$ incrementos atípicos num total de $n(n-1)/2$ incrementos, os quais são utilizados para estimar o variograma. Assim, para uma observação atípica, obtém-se uma proporção de 2 incrementos atípicos por cada n incrementos que são utilizados na estimação pontual do variograma. Logo, é necessário ter em atenção, que as observações atípicas se replicam no cálculo dos incrementos,

tendo um efeito maior do que à partida era de esperar.

Na Figura 6.1 procura-se ilustrar o efeito que a ampliação das observações atípicas (provocado pelo cálculo dos incrementos) tem sobre o estimador de Matheron. Esta figura foi construída com a mesma amostra que foi utilizada para obter as estimativas pontuais do semivariograma da Figura 2.1, substituindo apenas uma das observações. Em particular, substituiu-se uma observação da amostra do processo (que foi seleccio-

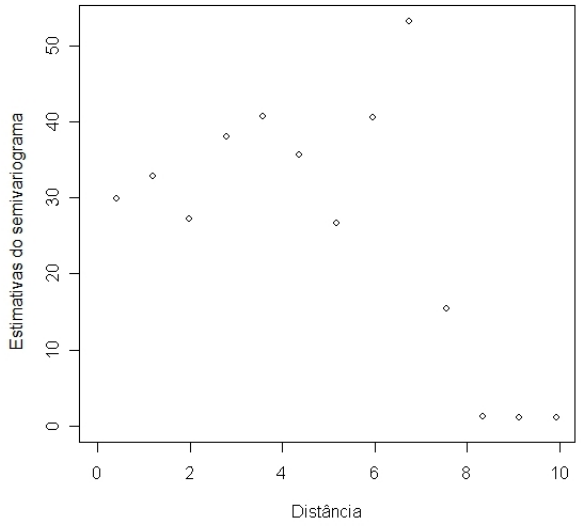

Figura 6.1: Estimativas pontuais do semivariograma calculadas com a mesma amostra usada na Figura 2.1, após a substituição de uma única observação.

nada aleatoriamente), por outra igual a cinquenta vezes o máximo da amostra.

Como se vê na Figura 6.1, a observação discordante alterou completamente todas as estimativas pontuais do semivariograma. A alteração foi tão forte que deixa de se notar o padrão de dependência evidenciado na Figura 2.1. Uma análise preliminar baseada na Figura 6.1, possivelmente, conduziria o analista a supor que as observações do processo tinham uma estrutura de dependência completamente diferente da que as gerou, ou que poderia nem haver interesse na estimação do variograma.

Os procedimentos que se utilizam tradicionalmente na segunda etapa da estimação

do variograma também não são robustos. De facto, os estimadores de mínimos quadrados (simples ou com a ponderação sugerida em Cressie (1993)) têm função de influência ilimitada e ponto de ruptura nulo (*vide* Hampel *et al.* (1986)). Portanto, se as estimativas pontuais do variograma estiverem contaminadas, o estimador de mínimos quadrados também vai reflectir essa contaminação nas estimativas dos parâmetros do modelo.

Contudo, não adianta usar um estimador robusto na segunda etapa da estimação, sem acautelar a estimação robusta na primeira etapa; quando o efeito ampliador das observações atípicas resulta na má estimação pontual do variograma, mesmo que seja usado um estimador robusto na segunda etapa, não será possível que ele capte a estrutura de dependência da amostra original. Portanto, o ponto de ruptura de todo o procedimento passa a ser nulo, uma vez que apenas uma observação atípica consegue destruir o resultado final da estimação do variograma. Por exemplo, na ilustração considerada na Figura 6.1, um estimador robusto iria produzir estimativas baseadas apenas nos pontos aí representados graficamente, os quais já não revelam a estrutura de dependência da amostra original.

Em face do que se acabou de referir, é de concluir que, quando se pretende um estimador robusto do variograma, é imprescindível que se utilize estimação robusta logo na primeira etapa de estimação.

Admitindo que é usado um estimador robusto na primeira fase da estimação, então, pelo próprio conceito de robustez, as estimativas pontuais não deverão ter sido muito afectadas por algum afastamento em relação às hipóteses assumidas no modelo. Deste modo, é possível abordar a segunda etapa de estimação do variograma, ou seja, a estimação dos parâmetros do modelo de variograma, como se todas as hipóteses inicialmente assumidas, se verificassem aproximadamente. Concluindo, é possível utilizar um procedimento não robusto na segunda etapa de estimação, sem prejudicar a robustez do método no seu todo. Procedendo desse modo, a robustez do estimador que se usa na primeira etapa, vai definir as características de robustez do processo no seu todo.

Apesar disso, a estimação dos parâmetros do modelo de variograma (segunda etapa) também pode ser feita através de um método robusto. De facto, podem-se utilizar, simultaneamente, métodos robustos na primeira e na segunda etapas de estimação do

variograma. Contudo, em modelos normais, quanto mais robusto for um estimador, menos eficiente ele se torna. Isto significa que, quando se utiliza um estimador robusto na primeira etapa da estimação, a utilização de um estimador robusto na segunda etapa, leva à perda de eficiência do estimador global; e geralmente não traz grande vantagem do ponto de vista da robustez. Consequentemente, na maior parte dos processos, o aumento da robustez do método da segunda etapa, não compensa a perda de eficiência do estimador global.

Concluindo: para obter um estimador do variograma simultaneamente robusto e com eficiência elevada, é de considerar um estimador pontual do variograma robusto, para resistir bem à quebra das hipóteses do modelo, e um método de estimação dos parâmetros do modelo de variograma o mais eficiente possível, pois a robustez na segunda etapa não é muito relevante.

Posto isto, seguidamente apresentam-se alguns estimadores pontuais do variograma que são robustos e que têm muito relevo na literatura.

6.2 Alguns estimadores pontuais robustos

Uma das primeiras propostas para tornar robusto o estimador pontual do variograma surgiu em Cressie e Hawkins (1980). A ideia consistiu em remover o quadrado dos incrementos presente no estimador de Matheron (*vide* ponto (2.1.1)), uma vez que este aumenta, significativamente, o impacto negativo que os incrementos atípicos provocam no estimador.

Se o processo $Z(\mathbf{s})$ for Gaussiano, então

$$\frac{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)}{\sqrt{2\gamma(\mathbf{s}_i - \mathbf{s}_j)}} \sim N(0,1);$$

donde, elevando ao quadrado, se obtém que

$$\frac{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2}{2\gamma(\mathbf{s}_i - \mathbf{s}_j)} \sim \chi^2_1. \tag{6.2.1}$$

Os autores verificaram que a raiz quarta de $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$, *i.e.* $|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}}$, é uma variável aleatória com distribuição aproximadamente normal, de média

$$\mathrm{E}[|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}}] \approx \frac{\Gamma(0.75)}{2\sqrt{\pi}} \times \gamma(\mathbf{s}_i - \mathbf{s}_j)^{1/4},$$

onde $\Gamma(x)$ representa a função Gama tradicional. Como consequência, também verificaram que o valor esperado de

$$\left(\frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}} \right)^4,$$

(onde $N(\mathbf{h})$ é o conjunto definido em (2.1.2)), é aproximadamente igual a

$$2\gamma(\mathbf{h}) \times \left(0.457 + \frac{0.494}{\#N(\mathbf{h})} + \frac{0.045}{(\#N(\mathbf{h}))^2} \right).$$

Deste resultado obtém-se um estimador do variograma empírico de expressão

$$2\bar{\gamma}(\mathbf{h}) = \frac{\left(\frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}} \right)^4}{0.457 + \frac{0.494}{\#N(\mathbf{h})} + \frac{0.045}{(\#N(\mathbf{h}))^2}}.$$

Contudo, observando a expressão anterior, é possível verificar que o termo $0.045/(\#N(\mathbf{h}))^2$ afecta muito pouco as estimativas, principalmente quando $\#N(\mathbf{h})$ é elevado. Por isso, Cressie e Hawkins (1980) decidiram definir o estimador

$$2\hat{\gamma}_{CH}(\mathbf{h}) = \frac{\left(\frac{1}{\#N(\mathbf{h})} \sum_{(\mathbf{s}_i,\mathbf{s}_j)\in N(\mathbf{h})} |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}} \right)^4}{0.457 + \frac{0.494}{\#N(\mathbf{h})}}. \tag{6.2.2}$$

Este estimador apresenta algumas boas propriedades:

- é aproximadamente centrado em processos Gaussianos;
- os termos $|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{\frac{1}{2}}$ são menos correlacionados do que os termos $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$ do estimador de Matheron (*vide* Hawkins (1981)).

A última propriedade traz vantagens quando se efectuam cálculos, para os quais é necessário assumir a independência dos incrementos (embora se saiba que ela não existe). Situações como essa ocorrem, por exemplo, na segunda fase de estimação do variograma, ao utilizar o estimador de mínimos quadrados ponderados, em substituição do estimador de mínimos quadrados generalizados.

Para além disso, o estimador de Cressie e Hawkins atenua o efeito de incrementos contaminados. De facto, os incrementos extremos são atenuados pelas raízes quadradas, o que não acontece com o estimador de Matheron. Como consequência, os incrementos atípicos têm um impacto substancialmente maior sobre o estimador de Matheron do

que sobre o estimador de Cressie e Hawkins. Este facto levou Cressie e Hawkins (1980) a apelidarem o seu estimador de robusto. No entanto, de facto, o estimador não é robusto, dado que tem função de influência ilimitada e ponto de ruptura nulo (*vide* Genton (1998a)).

Assim, de acordo com os critérios de robustez actualmente reconhecidos e referidos no **Capítulo 3**, nem o estimador de Matheron, nem o estimador de Cressie e Hawkins, são robustos.

Por outro lado, o estimador de Cressie e Hawkins não é muito eficiente – Genton (1998a) refere que a eficiência do estimador de Cressie e Hawkins sob modelos normais é de apenas 69.3%.

Uma maneira de melhorar a robustez dos estimadores acima referidos, consiste em substituir as médias que entram nas suas definições, pelas correspondentes medianas.

Uma vez que, sob as hipóteses usuais e tendo em atenção (6.2.1),

$$(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 \sim \chi^2_1 \times 2\gamma(\mathbf{s}_i - \mathbf{s}_j),$$

então o quantil amostral de ordem θ de $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$ estima $F^{-1}_{\chi^2_1}(\theta) \times 2\gamma(\mathbf{s}_i - \mathbf{s}_j)$, onde $F^{-1}_{\chi^2_1}(\theta)$ representa o quantil de ordem θ da distribuição do qui-quadrado com um grau de liberdade. Daqui se obtém um estimador pontual do variograma definido por

$$2\hat{\gamma}_Q(\mathbf{h}) = Q_\theta \left\{ (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h}) \right\} / F^{-1}_{\chi^2_1}(\theta),$$

onde $N(\mathbf{h})$ é o conjunto definido em (2.1.2) e $Q_\theta\{.\}$ denota o quantil empírico de ordem θ. No caso particular de $\theta = 1/2$, obtém-se o estimador baseado na mediana

$$2\hat{\gamma}_Q(\mathbf{h}) = \text{Mediana} \left\{ (Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2 : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h}) \right\} / 0.4549, \tag{6.2.3}$$

o qual resulta do estimador de Matheron, substituindo a média aritmética pela mediana.

Fazendo o mesmo raciocínio no estimador de Cressie e Hawkins e substituindo a média pela mediana, obtém-se

$$2\bar{\gamma}(\mathbf{h}) = \left[\text{Mediana} \left\{ |Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2} : (\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{h}) \right\} \right]^4 / 0.4549.$$

No entanto, tal como refere Cressie (1993), este estimador é equivalente ao (6.2.3). Na realidade, como a função $f(x) = x^{1/4}$ é monótona crescente para todo o $x > 0$, se o número de observações for elevado, tem-se que

$$\text{Mediana}\left\{(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2\right\} \approx \left[\text{Mediana}\left\{|Z(\mathbf{s}_i) - Z(\mathbf{s}_j)|^{1/2}\right\}\right]^4.$$

O estimador pontual do variograma $2\hat{\gamma}_Q$ apresentado em 6.2.3 que é baseado na mediana, é centrado, tem função de influência limitada e tem um ponto de ruptura de 50% na distribuição dos incrementos (máximo ponto de ruptura). Apesar de revelar boas propriedades no que diz respeito à robustez, este estimador tem como desvantagem, uma baixa eficiência, uma vez que a mediana amostral tem uma eficiência de apenas 63,7% em modelos normais.

Uma vez que os estimadores até agora referidos, ou não são robustos, ou têm baixa eficiência em modelos normais, Genton (1998a) sugeriu um estimador pontual do variograma robusto e com eficiência elevada. A ideia fundamental consistiu em estimar directamente a variância dos incrementos, utilizando um estimador de escala robusto e com boa eficiência em modelos normais. Assim, Genton utilizou o estimador Q_n de escala, proposto por Rousseeuw e Croux (1993), o qual foi apresentado na subsecção **3.3.4**.

Quando se aplica o estimador Q_n ao conjunto de incrementos separados por um vector $\mathbf{h}$, *i.e.* $N(\mathbf{h})$, está-se a estimar de forma robusta o desvio padrão dos incrementos, ou seja, está-se a estimar a raiz quadrada do variograma em $\mathbf{h}$. Assim, considere-se que

$$\begin{aligned}
&X_1 = Z(\mathbf{s}_1) - Z(\mathbf{s}_1 + \mathbf{h});\\
&X_2 = Z(\mathbf{s}_2) - Z(\mathbf{s}_2 + \mathbf{h});\\
&...\\
&X_{\#N(\mathbf{h})} = Z(\mathbf{s}_{\#N(\mathbf{h})}) - Z(\mathbf{s}_{\#N(\mathbf{h})} + \mathbf{h}).
\end{aligned}$$

De acordo com a definição do estimador Q_n apresentada em (3.3.6), este estimador toma a forma

$$Q_{\#N(\mathbf{h})} = c \times \{|Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \mathbf{h}) - (Z(\mathbf{s}_j) - Z(\mathbf{s}_j + \mathbf{h}))| : i < j\}_{(k)},$$

onde $k = \binom{[\#N(\mathbf{h})/2]+1}{2}$ e $c = 2.2191$.

Portanto, para estimar o variograma, ou seja, a variância dos incrementos, basta considerar o estimador

$$2\hat{\gamma}_G(\mathbf{h}) = Q^2_{\#N(\mathbf{h})}. \tag{6.2.4}$$

O estimador (6.2.4) é designado por estimador Q_n de Genton.

O estimador Q_n de Genton, contrariamente aos que foram abordados anteriormente, apresenta um enviesamento positivo em amostras com poucas observações, ou seja, quando $\#N(\mathbf{h})$ é pequeno. Por isso, se $\#N(\mathbf{h})$ for pequeno, deve-se utilizar um factor de correcção do enviesamento. Croux e Rousseeuw (1992) apresentaram os factores de correcção do enviesamento do estimador $2\hat{\gamma}_G(\mathbf{h})$ para $\#N(\mathbf{h}) < 40$.

Tal como refere Genton (1998a), este estimador do variograma é consistente, tem uma eficiência assintótica em modelos normais de 82%, um ponto de ruptura de 50% na distribuição dos incrementos e uma função de influência limitada, com $\gamma^*(Q_n; \Phi) = 2.069$ na distribuição normal standardizada. Repare-se que o ponto de ruptura apresentado, refere-se ao número de incrementos atípicos, e não ao número de observações atípicas do processo que podem existir na amostra original. Em Genton (2001) é apresentado um estudo de simulação, que indica que $2\hat{\gamma}_G(\mathbf{h})$ resiste até cerca de 30% de observações atípicas na amostra do processo $Z(\mathbf{s})$.

As propriedades enunciadas, fazem com que o estimador Q_n de Genton seja o estimador pontual do variograma que melhor consegue conciliar a robustez com a eficiência em modelos normais.

6.3 Uma população de variogramas empíricos

O conjunto de estimativas pontuais do variograma, que é obtido através da amostra do processo $Z(\mathbf{s})$, designa-se frequentemente por variograma empírico. No seguimento, utilizar-se-á essa designação, quando o conjunto de estimativas pontuais do variograma for encarado como um instrumento de análise preliminar de dados.

Nesta secção, apresenta-se um método de análise preliminar de dados, que permite fazer conjecturas sobre o variograma teórico que melhor modela um determinado conjunto de dados. Como foi visto no **Capítulo 2**, o variograma empírico é uma ferramenta crucial nessa escolha do modelo de variograma. Também é a partir do

variograma empírico que, posteriormente, se estimam os parâmetros desconhecidos do modelo de variograma escolhido.

Teoricamente, o variograma empírico é encarado como sendo único, ou seja, cada amostra concreta do processo $Z(\mathbf{s})$ dá origem a apenas um variograma empírico. Mas isto só acontece, quando cada ponto do variograma empírico é determinado exactamente pelas localizações que estão separadas pelo vector $\mathbf{h} \in \mathbb{R}^d$. Contudo, na grande maioria das aplicações, o variograma empírico não se obtém exactamente deste modo, dado que é frequente encontrarem-se vectores que separam poucos pares de localizações. Consequentemente, o variograma empírico é determinado à custa de regiões de tolerância dos vectores $\mathbf{h}$. Isto significa que, um ponto de um variograma empírico é determinado por todos os pares de localizações que estão separadas por vectores que, de algum modo, estão próximos de $\mathbf{h}$. Por exemplo, em processos de variograma isotrópico, a proximidade é medida em termos de $\|\mathbf{h}\|$, como foi referido na secção **2.1**.

A ideia que preside à utilização de regiões de tolerância, resume-se a ter em conta que, se o vector $\mathbf{s}_i - \mathbf{s}_j$ está próximo do vector $\mathbf{s}'_i - \mathbf{s}'_j$, então os valores do variograma nesses vectores, também devem ser próximos entre si, *i.e.*, $2\gamma(\mathbf{s}_i - \mathbf{s}_j) \approx 2\gamma(\mathbf{s}'_i - \mathbf{s}'_j)$. Assim, é adequado e conveniente agrupar os vectores próximos, para obter um dado ponto do variograma empírico.

Uma das situações em que se torna imprescindível a utilização de regiões de tolerância, ocorre quando as localizações da amostra estão dispostas de uma forma irregular. Nesse caso, é bastante frequente encontrarem-se vectores que separam apenas um par de localizações, ou seja, existe apenas um par $(\mathbf{s}_i, \mathbf{s}_j)$ tal que $\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}$. Por isso, o variograma empírico deve ser determinado através de regiões de tolerância.

Apesar das regiões de tolerância serem amplamente utilizadas na prática, elas geram alguma ambiguidade no variograma empírico. Mais precisamente, cada região de tolerância considerada, dá origem a um variograma empírico diferente. A Figura 6.2 representa esta ambiguidade. Nela podem-se observar quatro variogramas empíricos, que foram obtidos a partir da mesma amostra de um processo de variograma isotrópico. Todos os variogramas empíricos foram determinados pela expressão (2.1.1). A única diferença de representação para representação, reside nas regiões de tolerância consideradas. É possível verificar que, apesar dos quatro variogramas empíricos exibirem

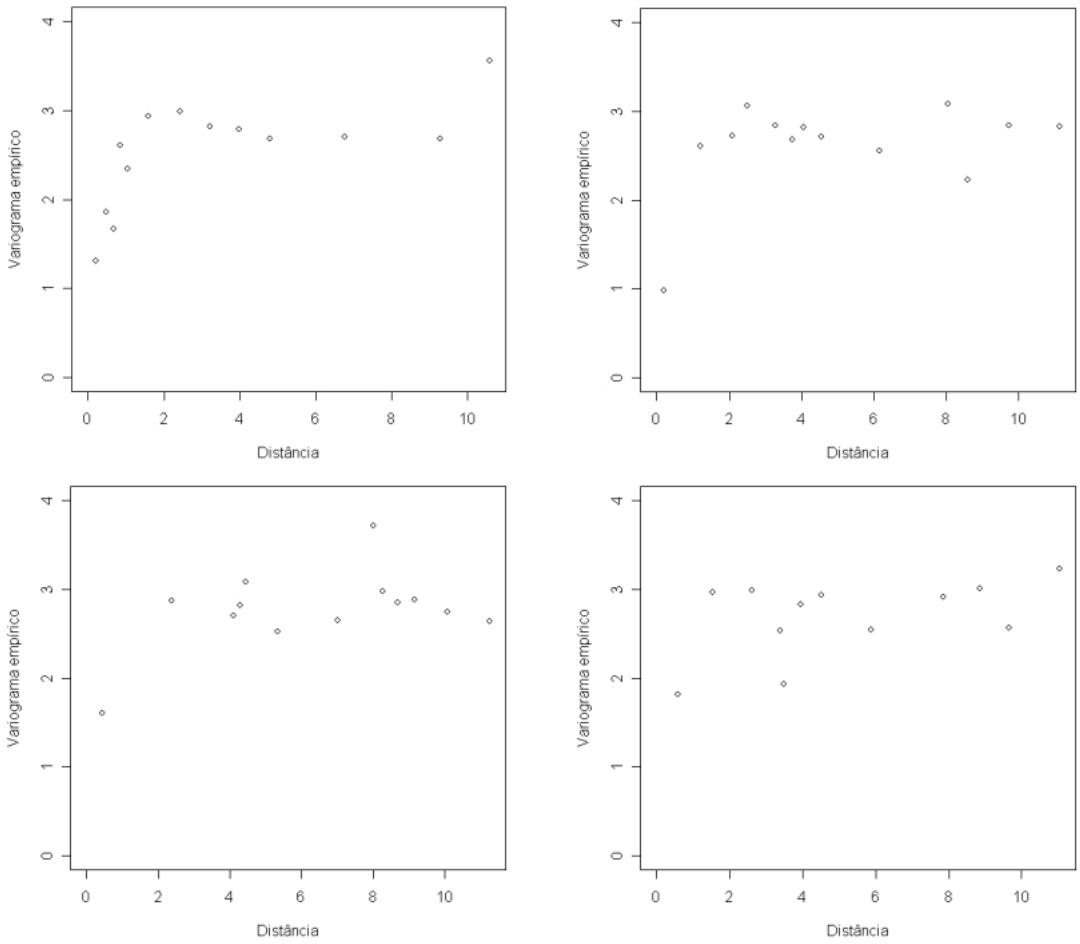

Figura 6.2: Quatro variogramas empíricos obtidos a partir da mesma amostra de um processo $Z(\mathbf{s})$ de variograma isotrópico, variando apenas as regiões de tolerância.

sensivelmente o mesmo tipo de padrão, existe uma variabilidade diferente em cada um deles – compare-se, por exemplo, o do canto superior esquerdo com o do canto inferior direito da figura.

Surge assim a motivação para considerar um conjunto de variogramas empíricos que se podem obter a partir de uma amostra fixa, em vez de ter em conta apenas um variograma empírico.

Considere-se assim o conjunto Ω de todos os variogramas empíricos distintos, que podem ser obtidos a partir de uma amostra $\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)\}$ do processo $Z(\mathbf{s})$, variando apenas as regiões de tolerância consideradas. O conjunto Ω pode ser encarado como uma população de variogramas empíricos da amostra $\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)\}$.

Tal como foi referido na secção **2.1**, ao construir as regiões de tolerância, há que ter em atenção alguns aspectos – as regiões não devem ser muito pequenas, para evitar que incluam poucos incrementos do processo; também não devem ser muito grandes, pois perde-se informação sobre a estrutura de dependência presente dentro de cada região, tornando o variograma empírico mais grosseiro.

O processo usual de escolha do modelo de variograma, é baseado apenas num único variograma empírico de Ω. Mas, é possível considerar uma amostra aleatória $(2\hat{\gamma}_1, ..., 2\hat{\gamma}_B)$ de variogramas empíricos da população Ω e, a partir dessa amostra, fazer conjecturas sobre o modelo de variograma adequado. Note-se que, no passado, esta abordagem teria sido inviável devido a dificuldades computacionais. No entanto, o desenvolvimento dos meios informáticos, criou condições para que, actualmente, se explore esta via com grande facilidade.

A Figura 6.3 ilustra a sugestão anteriormente apresentada. Nela pode-se visualizar o gráfico obtido representando, simultaneamente, cinquenta variogramas empíricos, calculados a partir da mesma amostra que foi utilizada para construir a Figura 6.2. Como se pode verificar, há vantagem em considerar uma amostra de variogramas empíricos, pois evidencia melhor a estrutura de dependência existente e facilita a posterior selecção do modelo de variograma.

Se o estimador usado para obter as estimativas pontuais do variograma for centrado, tal como é o estimador de Matheron, todos os variogramas empíricos pertencentes à população Ω têm valor esperado aproximadamente igual ao variograma do processo $Z(\mathbf{s})$, isto é,

$$\forall_{2\hat{\gamma}\in\Omega}\, \forall_{\mathbf{h}\in D_{2\hat{\gamma}}} \quad \mathrm{E}[2\hat{\gamma}(\mathbf{h})] \approx 2\gamma(\mathbf{h}),$$

onde $D_{2\hat{\gamma}}$ representa o conjunto dos vectores onde $2\hat{\gamma}$ foi determinado. Repare-se que, todos os variogramas empíricos da amostra aleatória $(2\hat{\gamma}_1, ..., 2\hat{\gamma}_B)$ são apenas aproximadamente centrados, uma vez que as regiões de tolerância podem introduzir algum enviesamento nas estimativas pontuais.

O método de reamostrar múltiplos variogramas empíricos a partir de uma única amostra, também é útil para obter uma aproximação inicial para os parâmetros do modelo de variograma em questão. Para tal, deve-se utilizar um estimador robusto, por motivos já analisados na secção **6.1**. Assim, na construção de cada variograma

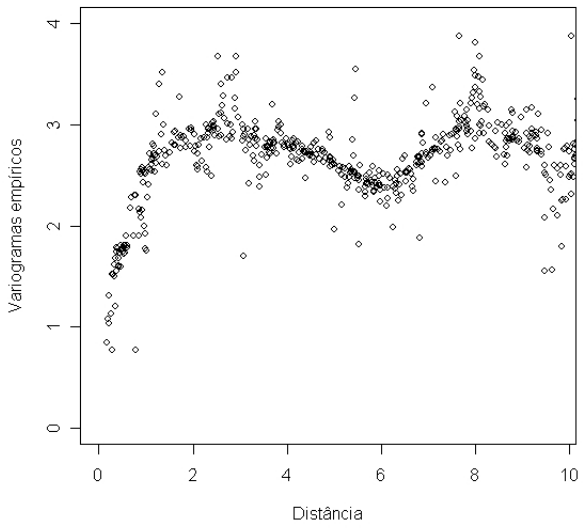

Figura 6.3: Cinquenta variogramas empíricos obtidos através de (2.1.1), a partir da mesma amostra da Figura 6.2.

empírico, recomenda-se, por exemplo, a utilização do estimador Q_n de Genton. As estimativas pontuais obtidas, revelam uma aproximação inicial resistente, que é útil para estimar os parâmetros do modelo de variograma.

Como se referiu, a obtenção de múltiplos variogramas empíricos facilita a escolha do modelo e a obtenção de estimativas iniciais para os parâmetros desse modelo. No entanto, por razões que se explicam no seguimento, esse conjunto de variogramas pontuais, não deve ser utilizado como ponto de partida para a segunda etapa da estimação.

De facto, caso se considerasse o conjunto formado por todos os pontos, obtidos com todos os diferentes variogramas empíricos, ficar-se-ia com uma nuvem de pontos, a qual era constituída por observações com uma estrutura de dependência demasiado complexa – por um lado, os variogramas empíricos são independentes entre si; mas por outro lado, cada variograma é constituído por observações correlacionadas. Desse modo, numa única representação, passariam a figurar observações correlacionadas e não correlacionadas, sem distinção. Assim, não haveria garantias de que os métodos a usar na segunda fase de estimação, gozassem de boas propriedades.

Pelos motivos expostos, o método apresentado na presente secção, não tem como objectivo a estimação dos parâmetros do modelo de variograma. A sua utilidade revela-se numa fase inicial de análise preliminar de dados, quando se pretende investigar qual o modelo de variograma que é mais adequado, ou quando se pretende encontrar uma aproximação inicial para os seus parâmetros.

6.4 O estimador de múltiplos variogramas

Nesta secção, apresenta-se um método de estimação do variograma que consegue conciliar boas propriedades de robustez com boa eficiência em modelos normais. A metodologia, que será designada por estimador de múltiplos variogramas, desenvolve-se em quatro etapas distintas. A primeira e a segunda etapas são semelhantes às etapas de estimação tradicional do variograma, isto é, consistem na estimação pontual do variograma (mas usando um método robusto), seguida da estimação dos parâmetros do modelo. Na terceira etapa, fazem-se variar os pontos onde o variograma é estimado, repetindo as etapas anteriores em cada caso; assim, obtém-se um conjunto de estimativas dos parâmetros do modelo de variograma (obtidas de forma robusta). Por fim, a última etapa consiste em determinar a estimativa final do variograma, com base no conjunto de estimativas do variograma já disponíveis. O nome "estimador de múltiplos variogramas"é motivado pela terceira etapa do processo.

6.4.1 Metodologia

Tendo em conta os argumentos apresentados na secção **6.1**, para que o resultado final da estimação do variograma seja robusto, deve-se utilizar um estimador robusto logo na primeira etapa de estimação do variograma, quando se determina o conjunto das estimativas pontuais, qualquer que seja o método a utilizar na segunda etapa de estimação. Sendo assim, para que o estimador global do variograma tenha boas propriedades, quer no que toca à robustez, quer no que diz respeito à eficiência em modelos normais, deve-se considerar um estimador pontual do variograma que seja simultaneamente robusto e com boa eficiência e, na segunda fase, um estimador dos parâmetros do modelo de variograma que seja o mais eficiente possível.

O estimador pontual do variograma que melhor consegue conciliar boas propriedades de robustez com boa eficiência em modelos normais é o estimador Q_n de Genton, como já foi referido na secção **6.2**. Sendo assim, considera-se que este é o estimador mais indicado para proceder à estimação pontual do variograma (primeira etapa de estimação).

Por outro lado, em Lahiri *et al.* (2002) foi demonstrado que o método dos mínimos quadrados generalizados (*GLS*) é o método mais eficiente para a estimação dos parâmetros do modelo de variograma, assumindo a dependência entre observações de um processo Gaussiano. No entanto, em termos práticos, o *GLS* é bastante difícil de implementar, mesmo quando se utiliza o estimador de Matheron na primeira etapa da estimação. As dificuldades surgem, porque a matriz de covariâncias do estimador pontual do variograma depende do próprio variograma que se está a estimar, e porque a forma dessa matriz é muito complicada, mesmo quando os modelos são normais. Isto torna impossível a utilização dos *GLS*, principalmente quando as estimativas pontuais do variograma são obtidas através de um estimador robusto. Genton (1998b) refere que a matriz de covariâncias do estimador Q_n é impossível de determinar analiticamente, mesmo quando as observações do processo são independentes. O método dos múltiplos variogramas vai permitir contornar a dificuldade de utilização dos *GLS*.

Uma das ideias fundamentais que estiveram na base da proposta do estimador de múltiplos variogramas foi sugerida pelo artigo de Lahiri *et al.* (2002). Nesse artigo, os autores mostram que, sob certas condições de regularidade, o estimador *GLS* tem a mesma eficiência assintótica do que o estimador de mínimos quadrados ponderados (*WLS*), ou do que o estimador de mínimos quadrados usuais (*OLS*), desde que a estimação seja efectuada utilizando um número de estimativas pontuais igual ao número de parâmetros que se pretendem estimar no modelo de variograma. Por isso, desde que se tenha um estimador consistente do variograma na primeira etapa de estimação, calculado num número de pontos igual ao número de parâmetros do modelo de variograma, então, à medida que o número de observações da amostra aumenta, a utilização do *OLS*, do *WLS* ou do *GLS*, na segunda etapa, conduz às mesmas propriedades assintóticas do estimador global, uma vez que aqueles estimadores são todos assintoticamente eficientes.

Contudo, para que o estimador global do variograma tenha boas propriedades, não basta considerar um número de estimativas pontuais (obtidas com um estimador robusto) igual ao número de parâmetros desconhecidos do modelo e, de seguida, estimar os parâmetros do modelo por *OLS*. É preciso ter em conta que, quando se utiliza este método sem as devidas precauções, podem-se obter resultados muito pouco precisos. Repare-se que o número de parâmetros desconhecidos do modelo de variograma é bastante pequeno, normalmente na ordem dos três parâmetros. Consequentemente, como o número de estimativas pontuais do variograma tem que ser igual ao número de parâmetros a estimar, vão existir muito poucas observações para estimar os parâmetros do modelo de variograma.

Uma das consequências imediatas da existência de poucas estimativas pontuais do variograma é a possível falta de identificabilidade dos parâmetros do modelo. Isto quer dizer que, se as estimativas pontuais do variograma não satisfizerem algumas condições de regularidade, a solução de mínimos quadrados pode não ser única. Por isso, é fundamental estabelecer condições sobre as estimativas pontuais do variograma, para que a solução de mínimos quadrados seja única. Este tema vai ser desenvolvido na subsecção **6.4.2**.

A possível existência de soluções múltiplas no método dos mínimos quadrados, não é o único problema que pode ocorrer. De facto, no procedimento tradicional de estimação, é usual considerarem-se mais observações do que parâmetros. Ao considerar tantas estimativas pontuais do variograma quanto os parâmetros, isso implica um aumento significativo da variabilidade dos estimadores de mínimos quadrados, principalmente quando a amostra do processo em estudo é pequena. Para além disso, quando há poucas estimativas pontuais do variograma, as estimativas de mínimos quadrados dependem muito dos vectores $\mathbf{h}$ onde se estimou pontualmente o variograma. A Figura 6.4 ilustra esta situação, num processo de variograma isotrópico. Pode-se verificar que as curvas estimadas dependem muito das normas de $\mathbf{h}$, para as quais se obtiveram as estimativas pontuais do variograma.

Uma maneira de reduzir a variabilidade do estimador de mínimos quadrados sem aumentar o número de estimativas pontuais, consiste em arranjar um processo que

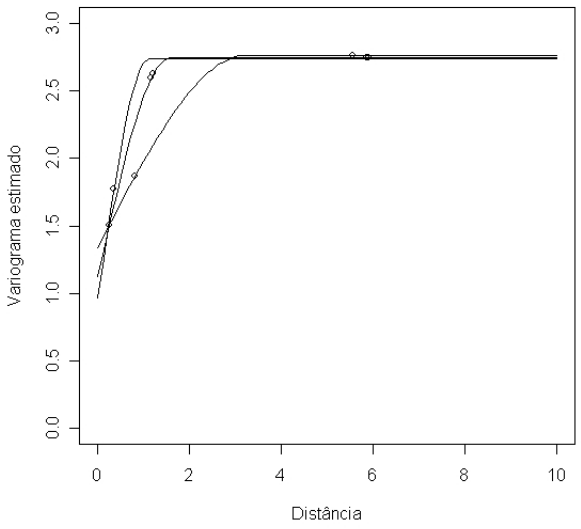

Figura 6.4: Diferentes estimativas do modelo de variograma em função da localização das estimativas pontuais.

disponibilize diversas estimativas dos parâmetros e, de entre elas, obter a estimativa final do variograma. É essa a ideia que se utiliza no desenvolvimento do método dos múltiplos variogramas.

Assim, na primeira etapa do método, estima-se pontualmente o variograma através do estimador Q_n de Genton. O número de estimativas pontuais do variograma deve ser igual ao número de parâmetros do modelo de variograma que se pretende estimar. Os vectores onde se estima o variograma pontualmente devem ser escolhidos por forma a não causarem problemas de identificabilidade (*vide* subsecção **6.4.2**).

Na segunda etapa, estimam-se os parâmetros do modelo de variograma através do estimador *OLS* que, no contexto do parágrafo anterior, é assintoticamente eficiente.

Na terceira etapa do procedimento de múltiplos variogramas, constrói-se um conjunto de estimativas dos parâmetros do modelo, estimadas a partir da mesma amostra, mas variando os locais onde se obtêm as estimativas pontuais do variograma. Note-se que o conjunto de estimativas dos parâmetros pode ser encarado como um conjunto de estimativas do variograma.

Por fim, na quarta etapa, partindo do conjunto das estimativas dos parâmetros do

modelo de variograma, usam-se medidas centrais dessas estimativas (neste caso com o recurso à mediana amostral), para definir a estimativa final do variograma.

Para concretizar as quatro fases do procedimento da estimação por múltiplos variogramas, considere-se o caso particular de um processo de variograma isotrópico com os três parâmetros usuais desconhecidos, nomeadamente, a amplitude, o patamar e o efeito de pepita.

Para que o método de *OLS* seja assintoticamente eficiente, na primeira etapa de estimação do variograma interessa obter três estimativas pontuais. Por outro lado, para que o método de *OLS* não tenha soluções múltiplas, as estimativas pontuais do variograma devem satisfazer algumas condições de regularidade (a ver detalhadamente na subsecção **6.4.2**). As estimativas pontuais do variograma são determinadas através do estimador Q_n de Genton, que é robusto e que tem boa eficiência em modelos normais. Na segunda etapa, estimam-se os parâmetros do modelo de variograma através do método de *OLS*. Nestas condições, o método de *OLS* é assintoticamente eficiente. Na terceira etapa, repetem-se as etapas anteriores variando os vectores onde se obtêm as estimativas pontuais do variograma. Assim, obtém-se um conjunto de estimativas $\{\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_B\}$ dos parâmetros do modelo de variograma, ou seja, obtém-se um conjunto de variogramas estimados a partir da mesma amostra de $Z(\mathbf{s})$. Para concluir, determina-se o variograma central do conjunto que foi obtido. Esse variograma central, que é encarado como a estimativa final do variograma, é obtido pelas medianas das estimativas dos parâmetros, $\{\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_B\}$.

Na subsecção **6.4.3** é apresentado um algoritmo de cálculo das estimativas.

6.4.2 O problema das soluções múltiplas

Ao longo da presente subsecção utiliza-se o semivariograma em vez do variograma, dado que o primeiro é vantajoso ao nível da apresentação de alguns resultados. No entanto, os resultados permanecem válidos para o variograma. Em todos os casos estudados, consideram-se apenas modelos de semivariograma isotrópicos, para facilitar a apresentação. Porém, é possível utilizar os mesmos resultados em semivariogramas anisotrópicos, fazendo a análise para cada direcção fixa considerada.

Como se referiu na subsecção anterior, a estimação dos parâmetros do modelo de

semivariograma a partir de poucas estimativas pontuais, pode resultar em problemas de identificabilidade e, consequentemente, na existência de soluções múltiplas ao determinar as estimativas pelo método dos mínimos quadrados usuais. A Figura 6.5 ilustra a facilidade com que as soluções múltiplas podem ocorrer, caso não se tomem as devidas precauções. Todos os semivariogramas representados são soluções de *OLS*, obtidas com as mesmas estimativas pontuais do semivariograma.

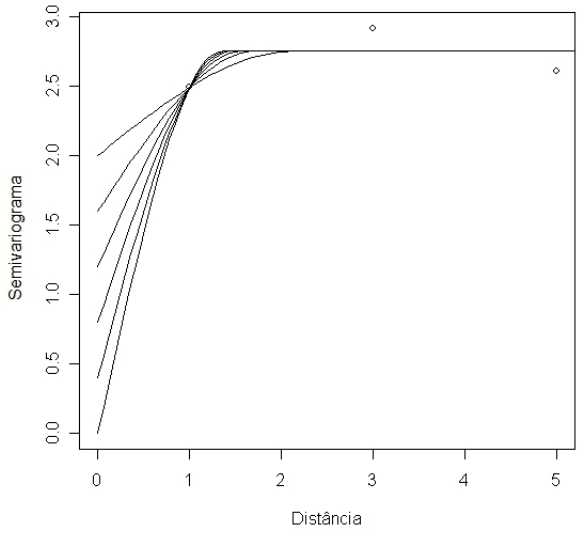

Figura 6.5: Semivariograma pontual constituído por três estimativas e múltiplas soluções obtidas por *OLS*.

A situação retratada na Figura 6.5 não é um caso raro e mostra bem como a presença de soluções múltiplas pode levar a maus resultados. Sendo assim, quando existem poucas estimativas pontuais do semivariograma, é fundamental estabelecer condições que garantam a unicidade de solução, na estimação dos parâmetros do modelo de semivariograma. Para além disso, os problemas de multiplicidade de solução, também se colocam no caso geral em que existe um número maior de estimativas pontuais. O estudo que se passa a apresentar foi desenvolvido para o caso limite de três estimativas pontuais do semivariograma, uma vez que são essas as condições utilizadas no estimador de múltiplos variogramas. No entanto, os resultados encontrados são aplicáveis ao estimador tradicional, com semivariograma empírico constituído por mais do que três

estimativas pontuais.

O problema das soluções múltiplas será abordado em três situações distintas – em primeiro lugar, estudam-se os processos geoestatísticos estacionários de segunda ordem que têm amplitude (*vide* subsecção **1.3.1**); depois, consideram-se os processos geoestatísticos estacionários de segunda ordem que não têm amplitude e que, por isso, só têm amplitude prática; por último, investigam-se os processos geoestatísticos que são apenas intrinsecamente estacionários e que, portanto, não têm amplitude nem amplitude prática.

Assim, considere-se um processo estacionário de segunda ordem com amplitude. Uma vez que existe amplitude, a partir desse valor, o semivariograma é uma função constante que assume um valor igual ao patamar. Para processos com as características referidas, os teoremas **6.4.1** e **6.4.2** que se apresentam seguidamente, estabelecem condições de existência de soluções múltiplas na estimação da amplitude, do efeito de pepita e do patamar.

Teorema 6.4.1. Seja $Z(\mathbf{s})$ um processo geoestatístico estacionário de segunda ordem com semivariograma $\gamma_0(h)$, pertencente a uma família $\mathfrak{F}$ de semivariogramas isotrópicos que são contínuos em $\mathbb{R}^+$, tal que

$$\begin{aligned}\mathfrak{F} &= \{\gamma(\|\mathbf{h}\|\,;\tau^2,\sigma^2,\phi) : \mathbf{h}\in\mathbb{R}^d \wedge \tau\in\mathbb{R}_0^+ \wedge \sigma,\phi\in\mathbb{R}^+\}\\ &= \{\gamma(h;\tau^2,\sigma^2,\phi) : h,\tau\in\mathbb{R}_0^+ \wedge \sigma,\phi\in\mathbb{R}^+\},\end{aligned}$$

onde $\tau^2, \tau^2+\sigma^2$ e ϕ representam, respectivamente, o efeito de pepita, o patamar e a amplitude dos semivariogramas. Sejam τ_0^2, σ_0^2 e ϕ_0 os valores dos parâmetros de $\gamma_0(h)$, isto é, $\gamma_0(h) = \gamma(h;\tau_0^2,\sigma_0^2,\phi_0)$, e h_1, h_2 e h_3 quaisquer três números reais positivos distintos.

Se existe, no máximo, um h_i $(i = 1,2,3)$ para o qual $h_i < \phi_0$, então existe mais do que um semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_1,\gamma_0(h_1)), (h_2,\gamma_0(h_2))$ e $(h_3,\gamma_0(h_3))$.

Demonstração: Como todos os semivariogramas de $\mathfrak{F}$ são semivariogramas de processos estacionários de segunda ordem, então qualquer processo com semivariograma $\gamma(h;\tau^2,\sigma^2,\phi)\in\mathfrak{F}$ possui um covariograma $C(h;\sigma^2,\phi)$, o qual não depende de τ^2. Em

consequência da equação (1.3.4), tem-se que

$$\forall_{h\in\mathbb{R}^+} \quad \gamma(h;\tau^2,\sigma^2,\phi) = \tau^2 + \sigma^2 - C(h;\sigma^2,\phi).$$

Mas, se $\gamma(h;\tau^2,\sigma^2,\phi)$ pertence a $\mathfrak{F}$, então ele é contínuo em $\mathbb{R}^+$ e tem amplitude $\phi \in \mathbb{R}^+$. Por definição de amplitude, para $h \geq \phi$, tem-se que $\gamma(h;\tau^2,\sigma^2,\phi) = \tau^2 + \sigma^2$. Deste modo, qualquer $\gamma(h;\tau^2,\sigma^2,\phi) \in \mathfrak{F}$ pode representar-se na forma

$$\gamma(h;\tau^2,\sigma^2,\phi) = \begin{cases} 0 & se \quad h = 0 \\ \tau^2 + \sigma^2 - C(h;\sigma^2,\phi) & se \quad 0 < h < \phi \\ \tau^2 + \sigma^2 & se \quad h \geq \phi \end{cases} . \tag{6.4.1}$$

Em particular, o semivariograma $\gamma_0(h)$ pertence a $\mathfrak{F}$ e pode ser escrito na forma (6.4.1), para os valores dos parâmetros τ_0^2, σ_0^2 e ϕ_0 que o definem.

Considerem-se agora $h_1, h_2, h_3 \in \mathbb{R}^+$ distintos, tais que existe, no máximo, um i, para o qual $h_i < \phi_0$ $(i = 1, 2, 3)$. Note-se que esta hipótese inclui dois casos: ou nenhum dos números h_i é menor do que ϕ_0, ou existe apenas um deles menor do que ϕ_0. Os dois casos são tratados separadamente.

Caso 1 $(\phi_0 \leq h_1 < h_2 < h_3)$:

Como nenhum dos h_i é inferior à amplitude ϕ_0, então, nesses pontos, o semivariograma $\gamma_0(h_i)$ $(i = 1, 2, 3)$ é determinado através do terceiro ramo de (6.4.1). Deste modo, $\gamma_0(h_1) = \gamma_0(h_2) = \gamma_0(h_3) = \tau_0^2 + \sigma_0^2$, ou seja, o semivariograma do processo passa pelos pontos $(h_1, \tau_0^2 + \sigma_0^2)$, $(h_2, \tau_0^2 + \sigma_0^2)$ e $(h_3, \tau_0^2 + \sigma_0^2)$.

Seja $\mathfrak{F}_1$ o subconjunto de $\mathfrak{F}$ constituído pelos semivariogramas com amplitude menor ou igual do que h_1, isto é,

$$\mathfrak{F}_1 = \{\gamma(h;\tau^2,\sigma^2,\phi) : h, \tau \in \mathbb{R}_0^+ \wedge \sigma \in \mathbb{R}^+ \wedge \phi \leq h_1\}.$$

Qualquer elemento de $\mathfrak{F}_1$ calculado no ponto h_i é obtido pelo terceiro ramo de (6.4.1) e assim, $\gamma(h_i;\tau^2,\sigma^2,\phi) = \tau^2 + \sigma^2$, para todo o $i = 1, 2, 3$.

Os semivariogramas de $\mathfrak{F}_1$ passam pelos pontos $(h_1, \gamma_0(h_1))$, $(h_2, \gamma_0(h_2))$ e $(h_3, \gamma_0(h_3))$, se e só se são solução do sistema de equações

$$\begin{cases} \gamma(h_1;\tau^2,\sigma^2,\phi) = \gamma_0(h_1) \\ \gamma(h_2;\tau^2,\sigma^2,\phi) = \gamma_0(h_2) \\ \gamma(h_3;\tau^2,\sigma^2,\phi) = \gamma_0(h_3) \end{cases} \Leftrightarrow \begin{cases} \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \end{cases} \Leftrightarrow \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2.$$

O sistema anterior resulta apenas numa equação com duas incógnitas, o τ^2 e o σ^2. Portanto, o sistema é possível e indeterminado. Assim, existe mais do que um semivariograma de $\mathfrak{F}_1$ que passa pelos pontos $(h_1, \gamma_0(h_1))$, $(h_2, \gamma_0(h_2))$ e $(h_3, \gamma_0(h_3))$. Como $\mathfrak{F}_1 \subset \mathfrak{F}$, então também existe mais do que um semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_i, \gamma_0(h_i))$, para $i = 1, 2, 3$. Mais precisamente, todos os semivariogramas de $\mathfrak{F}$ que verificam as condições $\phi \leq h_1$ e $\tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2$, passam pelos pontos pretendidos. Fica assim verificado o teorema para o primeiro caso considerado.

Caso 2 ($h_1 < \phi_0 \leq h_2 < h_3$):

Como $h_1 < \phi_0$, então $\gamma_0(h_1)$ é determinado através do segundo ramo de (6.4.1). Logo $\gamma_0(h_1) = \tau_0^2 + \sigma_0^2 - C(h_1; \sigma_0^2, \phi_0)$. Por outro lado, como h_2 e h_3 não são inferiores a ϕ_0, então, em qualquer um deles, $\gamma_0(h_i)$ é determinado através do terceiro ramo de (6.4.1). Assim $\gamma_0(h_2) = \gamma_0(h_3) = \tau_0^2 + \sigma_0^2$. Deste modo, o semivariograma $\gamma_0(h)$ passa pelos pontos $(h_1, \tau_0^2 + \sigma_0^2 - C(h_1; \sigma_0^2, \phi_0))$, $(h_2, \tau_0^2 + \sigma_0^2)$ e $(h_3, \tau_0^2 + \sigma_0^2)$.

Considere-se agora o subconjunto da família $\mathfrak{F}$ definido por

$$\mathfrak{F}_2 = \{\gamma(h; \tau^2, \sigma^2, \phi) : h, \tau \in \mathbb{R}_0^+ \wedge \sigma \in \mathbb{R}^+ \wedge h_1 < \phi \leq h_2\}.$$

Como h_1 é inferior a ϕ, então, qualquer elemento de $\mathfrak{F}_2$ calculado no ponto h_1, vem determinado através do segundo ramo de (6.4.1) e, assim, $\gamma(h_1; \tau^2, \sigma^2, \phi) = \tau^2 + \sigma^2 - C(h_1; \sigma^2, \phi)$. Por outro lado, como h_2 e h_3 não são inferiores a ϕ, $\gamma(h_i; \tau^2, \sigma^2, \phi)$ é determinado através do terceiro ramo de (6.4.1), para $i = 2, 3$. Ou seja, $\gamma(h_2) = \gamma(h_3) = \tau^2 + \sigma^2$.

Os semivariogramas $\gamma(h; \tau^2, \sigma^2, \phi) \in \mathfrak{F}_2$ passam pelos pontos $(h_1, \gamma_0(h_1))$, $(h_2, \gamma_0(h_2))$ e $(h_3, \gamma_0(h_3))$, se e só se são solução do sistema de equações

$$\begin{cases} \gamma(h_1; \tau^2, \sigma^2, \phi) = \gamma_0(h_1) \\ \gamma(h_2; \tau^2, \sigma^2, \phi) = \gamma_0(h_2) \\ \gamma(h_3; \tau^2, \sigma^2, \phi) = \gamma_0(h_3) \end{cases} \Leftrightarrow \begin{cases} \tau^2 + \sigma^2 - C(h_1; \sigma^2, \phi) = \tau_0^2 + \sigma_0^2 - C(h_1; \sigma_0^2, \phi_0) \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \end{cases}$$

$$\Leftrightarrow \begin{cases} C(h_1; \sigma^2, \phi) = C(h_1; \sigma_0^2, \phi_0) \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \end{cases}.$$

O sistema anterior resulta em duas equações a três incógnitas (τ^2, σ^2 e ϕ). Portanto, o sistema é possível e indeterminado e existe mais do que um semivariograma de $\mathfrak{F}_2$ que

passa pelos pontos $(h_1, \gamma_0(h_1)), (h_2, \gamma_0(h_2))$ e $(h_3, \gamma_0(h_3))$. Como $\mathfrak{F}_2 \subset \mathfrak{F}$, isso implica que existe mais do que um semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_i, \gamma_0(h_i))$, para $i = 1, 2, 3$. Mais precisamente, todos os semivariogramas de $\mathfrak{F}$ que verificam as condições $h_1 < \phi \leq h_2$, $C(h_1; \sigma_0^2, \phi_0) = C(h_1; \sigma^2, \phi)$ e $\tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2$, passam pelos pontos pretendidos. Fica assim demonstrado o teorema para o segundo caso considerado.

□

Teorema 6.4.2. Sejam $(h_i, \hat{\gamma}(h_i))$, para $i = 1, 2, 3$, as estimativas pontuais do semivariograma, obtidas a partir de um estimador consistente. Nas condições do **Teorema 6.4.1**, e para uma dimensão amostral suficientemente grande, existe mais do que um elemento de $\mathfrak{F}$ que é solução do problema de mínimos quadrados (OLS).

Demonstração: Seja $\gamma(h; \hat{\tau}^2, \hat{\sigma}^2, \hat{\phi})$ uma solução de mínimos quadrados obtida através das três estimativas pontuais do semivariograma $(h_i, \hat{\gamma}(h_i)), i = 1, 2, 3$. Deste modo,

$$(\hat{\tau}^2, \hat{\sigma}^2, \hat{\phi}) \in \arg\min_{\tau^2, \sigma^2, \phi} \sum_{i=1}^{3} \left(\gamma(h_i; \tau^2, \sigma^2, \phi) - \hat{\gamma}(h_i)\right)^2,$$

onde $\gamma(h; \tau^2, \sigma^2, \phi) \in \mathfrak{F}$ e $\hat{\gamma}(h_i)$ é o estimador pontual do semivariograma determinado em h_i, o qual é consistente para estimar $\gamma_0(h_i)$.

Como $\hat{\gamma}(h_i)$ é consistente, tem-se que

$$\hat{\gamma}(h_i) \xrightarrow{\mathcal{P}} \gamma_0(h_i),$$

para $i = 1, 2, 3,$, onde $\mathcal{P}$ significa convergência em probabilidade. Como a dimensão da amostra é suficientemente grande, qualquer solução de mínimos quadrados se aproxima do verdadeiro semivariograma nos pontos h_i, isto é, verifica-se que $\gamma(h_i; \hat{\tau}^2, \hat{\sigma}^2, \hat{\phi}) \approx \hat{\gamma}(h_i) \approx \gamma_0(h_i)$ $(i = 1, 2, 3)$.

Como consequência, existe uma curva de semivariograma que está próxima de $\gamma_0(h)$, para qualquer $h \in \mathbb{R}_0^+$, e que é solução de mínimos quadrados. Denote-se essa curva por $\gamma_1(h) = \gamma(h; \tau_1^2, \sigma_1^2, \phi_1)$. Como $\gamma_1(h) \approx \gamma_0(h)$, para qualquer h, os parâmetros que definem cada uma das curvas γ_0 e γ_1, têm valores semelhantes, e a consistência do estimador garante que, à medida que a dimensão da amostra aumenta, os valores desses parâmetros serão cada vez mais próximos. Logo, por um lado, tem-se que $\phi_0 \approx \phi_1$; por

outro lado, dada a relação entre os h_i e ϕ_0, tem-se que existe no máximo um $h_i < \phi_1$, para $i = 1, 2, 3$.

Portanto, a curva de semivariograma $\gamma_1(h)$, que é uma solução de mínimos quadrados, verifica as condições impostas a $\gamma_0(h)$ no teorema anterior. Sendo assim, esse teorema garante que existe mais do que um semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_i, \gamma_1(h_i))$, para $i = 1, 2, 3$, os quais também são soluções de mínimos quadrados.

□

Note-se que o **Teorema 6.4.2** só é válido quando o estimador pontual do semivariograma é consistente. Por isso, o resultado é aplicável em amostras de grande dimensão. Se a amostra for de pequena dimensão, a consistência do estimador já não garante que existe um variograma estimado próximo do verdadeiro variograma. No entanto, as conclusões do teorema ainda são aplicáveis, caso exista uma solução de mínimos quadrados próxima do verdadeiro variograma, reunindo as condições que, em grandes amostras, são asseguradas pela consistência do estimador usado na primeira etapa de estimação. Então, qualquer que seja a dimensão da amostra, conclui-se que, para estimar um semivariograma $\gamma_0(h)$ com amplitude, não devem existir menos do que dois pontos h_i, para $i = 1, 2, 3$, inferiores à amplitude ϕ_0 do semivariograma real.

O corolário seguinte particulariza a aplicação do **Teorema 6.4.2** para alguns modelos de semivariograma específicos.

Corolário 6.4.3. Considerem-se as famílias de semivariogramas de modelos de tenda, circular e esférico caracterizadas pelas expressões (1.3.5) – (1.3.7). Nas condições do **Teorema 6.4.2**, o método de mínimos quadrados (OLS) aplicado às estimativas pontuais do semivariograma $(h_i, \hat{\gamma}(h_i))$, para $i = 1, 2, 3$, tem mais do que uma solução.

Demonstração: Decorre imediatamente do **Teorema 6.4.2**, pelo facto das famílias de semivariogramas consideradas terem amplitude $\phi \in \mathbb{R}^+$.

□

Como exemplo do resultado anterior, considere-se um processo geoestatístico com modelo de semivariograma esférico isotrópico, com os três parâmetros (amplitude, efeito de pepita e patamar) desconhecidos. Tal como refere Soares (2000), o modelo esférico

é um dos modelos que é mais utilizado em Geoestatística. Considere-se ainda um estimador pontual do semivariograma que seja consistente, como por exemplo, o estimador Q_n de Genton. Como existem três parâmetros a estimar, então o método dos múltiplos variogramas supõe que se utilizem, exactamente, três estimativas pontuais do semivariograma. Supondo um modelo esférico, o **Corolário 6.4.3** assegura a existência de soluções múltiplas quando existe, no máximo, uma estimativa pontual do semivariograma com norma do vector **h** inferior à amplitude do processo considerado. A Figura 6.6 ilustra os dois casos de soluções múltiplas no modelo esférico – na figura da esquerda, as soluções múltiplas resultam do facto de existir apenas uma estimativa pontual com norma de **h** inferior à amplitude; à direita, as soluções múltiplas resultam do facto de não existir nenhuma estimativa pontual com norma de **h** inferior à amplitude. Todas as curvas da figura passam pelas três estimativas pontuais do semivariograma, logo são soluções do método de *OLS*. Sendo assim, em ambos os casos representados, quer a amplitude, quer o efeito de pepita, têm mais do que uma solução.

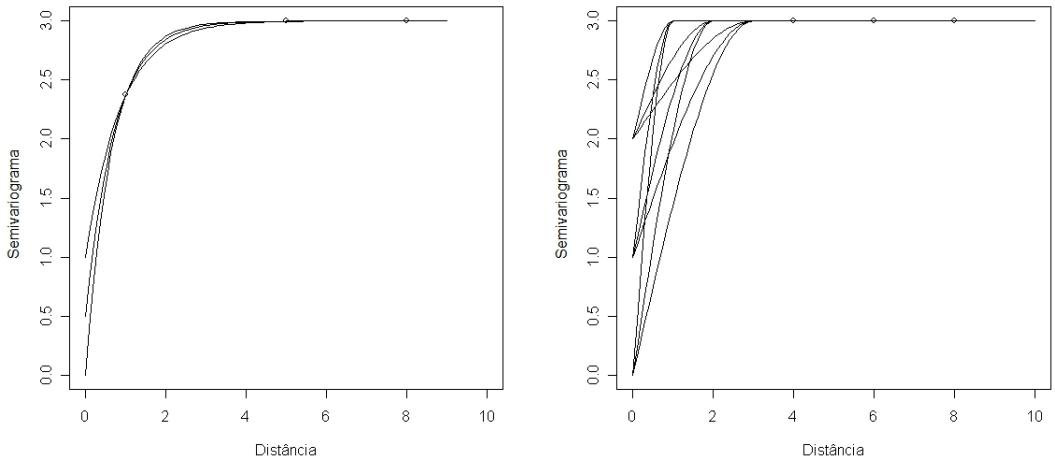

Figura 6.6: Ilustração de dois casos típicos onde existem soluções múltiplas na estimação dos parâmetros do modelo de semivariograma esférico.

No contexto de processos estacionários de segunda ordem com amplitude, resta investigar se existem condições que assegurem que o método dos mínimos quadrados usuais tenha solução única. De facto, os teoremas seguintes mostram que é possível

estabelecer tais condições, em função da posição relativa entre as estimativas pontuais do semivariograma e a amplitude.

Teorema 6.4.4. Seja $Z(\mathbf{s})$ um processo geoestatístico estacionário de segunda ordem, cujo semivariograma $\gamma_0(h)$ pertence à família $\mathfrak{F}$ definida no enunciado do **Teorema 6.4.1**. Considere-se que $\gamma_0(h;\tau^2,\sigma^2,\phi) = \gamma(h;\tau_0^2,\sigma_0^2,\phi_0)$. Seja $C(h;\sigma^2,\phi)$ a função covariograma correspondente ao elemento $\gamma(h;\tau^2,\sigma^2,\phi)$ de $\mathfrak{F}$, e h_1, h_2 e h_3, quaisquer três números reais positivos, tais que $h_1 < h_2 < \phi_0 \leq h_3$. Se existe apenas um par (σ^2,ϕ) para o qual

$$C(h_i;\sigma^2,\phi) = C(h_i;\sigma_0^2,\phi_0), \text{ quando } i = 1,2, \tag{6.4.2}$$

então existe um e um só semivariograma pertencente a $\mathfrak{F}$, que passa pelos pontos $(h_1,\gamma_0(h_1)), (h_2,\gamma_0(h_2))$ e $(h_3,\gamma_0(h_3))$, *i.e.*, existe apenas um valor de τ^2 para o qual $\gamma(h;\tau^2,\sigma^2,\phi)$ passa por todos os pontos $(h_i,\gamma_0(h_i))$, para $i = 1,2,3$.

Demonstração: Pretende-se verificar que, dadas as condições do teorema, o sistema

$$\begin{cases} \gamma(h_1;\tau^2,\sigma^2,\phi) = \gamma_0(h_1) \\ \gamma(h_2;\tau^2,\sigma^2,\phi) = \gamma_0(h_2) \\ \gamma(h_3;\tau^2,\sigma^2,\phi) = \gamma_0(h_3) \end{cases}$$

é possível e determinado.

Como o processo $Z(\mathbf{s})$ é estacionário de segunda ordem, para qualquer $\gamma(h;\tau^2,\sigma^2,\phi)$ $\in \mathfrak{F}$, tem-se que

$$\forall_{\tau\in\mathbb{R}_0^+} \quad \forall_{h,\sigma,\phi\in\mathbb{R}^+} \quad \gamma(h;\tau^2,\sigma^2,\phi) = \tau^2 + \sigma^2 - C(h;\sigma^2,\phi).$$

Então, o semivariograma $\gamma(h;\tau^2,\sigma^2,\phi)$, que depende de três parâmetros, pode ser escrito à custa do covariograma $C(h;\sigma^2,\phi)$, que depende apenas de σ^2 e de ϕ.

Como qualquer semivariograma $\gamma(h;\tau^2,\sigma^2,\phi) \in \mathfrak{F}$ tem amplitude $\phi \in \mathbb{R}^+$, então $\gamma(h;\tau^2,\sigma^2,\phi)$ pode ser escrito na forma (6.4.1). Em particular, $\gamma_0(h)$ também é determinado por (6.4.1), para os valores τ_0^2, σ_0^2 e ϕ_0 que o definem.

Considerem-se agora quaisquer três números reais h_1, h_2 e h_3, distintos, tais que $h_1 < h_2 < \phi_0 \leq h_3$. Por hipótese, o sistema de equações (6.4.2) tem solução única, o que implica que essa solução seja $(\sigma^2,\phi) = (\sigma_0^2,\phi_0)$.

Como $h_3 \geq \phi_0$, então o semivariograma γ_0 em h_3 coincide com o patamar, ou seja, $\gamma_0(h_3) = \tau_0^2 + \sigma_0^2$. Por outro lado, como pelo sistema (6.4.2) se sabe que $\phi = \phi_0$, então $h_3 \geq \phi$. Logo (6.4.1) implica que $\gamma(h_3) = \tau^2 + \sigma^2$.

Assim, o sistema anterior é equivalente a

$$\begin{cases} \tau^2 + \sigma^2 - C(h_1; \sigma^2, \phi)) = \tau_0^2 + \sigma_0^2 - C(h_1; \sigma_0^2, \phi_0)) \\ \tau^2 + \sigma^2 - C(h_2; \sigma^2, \phi)) = \tau_0^2 + \sigma_0^2 - C(h_2; \sigma_0^2, \phi_0)) \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \end{cases},$$

que se simplifica em

$$\begin{cases} C(h_1; \sigma^2, \phi)) = C(h_1; \sigma_0^2, \phi_0)) \\ C(h_2; \sigma^2, \phi)) = C(h_2; \sigma_0^2, \phi_0)) \\ \tau^2 + \sigma^2 = \tau_0^2 + \sigma_0^2 \end{cases};$$

mas, como o sistema formado pelas duas equações (6.4.2) é possível e determinado, então este último sistema tem solução única determinada por

$$\begin{cases} \phi = \phi_0 \\ \sigma^2 = \sigma_0^2 \\ \tau^2 = \tau_0^2 \end{cases}.$$

Portanto, existe um e um só semivariograma pertencente a $\mathfrak{F}$ que passa pelos pontos $(h_1, \gamma_0(h_1)), (h_2, \gamma_0(h_2))$ e $(h_3, \gamma_0(h_3))$, coincidente com o verdadeiro semivariograma, $\gamma_0(h)$.

□

O teorema que se segue é idêntico ao **Teorema 6.4.2** – a diferença entre ambos reside no facto de, no **Teorema 6.4.2**, serem estabelecidas condições suficientes para a existência de múltiplas soluções de *OLS*, enquanto que o **Teorema 6.4.5** vai enunciar condições suficientes para a unicidade de solução. Note-se que não existe complementaridade entre os dois resultados, pois em nenhum dos casos se apresentam condições necessárias e suficientes.

Teorema 6.4.5. Sejam $(h_i, \hat{\gamma}(h_i))$, para $i = 1, 2, 3$, as estimativas pontuais do semivariograma, obtidas a partir de um estimador consistente. Nas condições do **Teorema 6.4.4**, e para uma dimensão amostral suficientemente grande, existe um e um só elemento de $\mathfrak{F}$ que é solução do problema de mínimos quadrados (*OLS*).

Demonstração: Seja $\gamma(h;\hat{\tau}^2,\hat{\sigma}^2,\hat{\phi})$ uma solução de mínimos quadrados obtida através das três estimativas pontuais do semivariograma $(h_i,\hat{\gamma}(h_i))$, para $i=1,2,3$.

Como $\hat{\gamma}(h_i)$ é consistente, fazendo um raciocínio análogo ao da demonstração do **Teorema 6.4.2**, é possível concluir que, quando a dimensão da amostra é suficientemente grande, todas as soluções de mínimos quadrados verificam $\gamma(h_i;\hat{\tau}^2,\hat{\sigma}^2,\hat{\phi}) \approx \hat{\gamma}(h_i) \approx \gamma_0(h_i)$, para $i=1,2,3$. Assim, quando n tende para infinito, qualquer solução de mínimos quadrados é uma curva que passa pelos pontos $\gamma_0(h_i)$ $(i=1,2,3)$. Em particular, existe uma curva $\gamma_1(h)=\gamma(h;\tau_1^2,\sigma_1^2,\phi_1)$ que é solução de mínimos quadrados e que satisfaz $\gamma_1(h)\approx\gamma_0(h)$, para todo o h.

Consequentemente, os parâmetros que definem cada uma das curvas $\gamma_0(h)$ e $\gamma_1(h)$, têm valores semelhantes e, mais uma vez, a consistência do estimador garante que, à medida que a dimensão da amostra aumenta, os valores dos parâmetros de cada curva serão cada vez mais próximos. Logo, por um lado, tem-se que $\phi_0\approx\phi_1$; por outro lado, dada a relação entre os h_i e ϕ_0, tem-se que $h_1<h_2<\phi_1\leq h_3$.

Portanto, $\gamma_1(h)$ verifica as condições impostas a $\gamma_0(h)$ no Teorema 6.4.4. Logo, esse teorema garante que $\gamma_1(h)$ é o único semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_i,\gamma_1(h_i))$, para $i=1,2,3$, o qual é a única solução de mínimos quadrados.

□

Os três corolários seguintes apresentam a aplicação do resultado obtido no **Teorema 6.4.5**, particularizando o modelo de semivariograma aos modelos de tenda, circular e esférico.

Corolário 6.4.6. Seja $Z(\mathbf{s})$ um processo com semivariograma isotrópico $\gamma_0(h)=\gamma(h;\tau_0^2,\sigma_0^2,\phi_0)$ definido pelo modelo de tenda, e sejam $(h_i,\hat{\gamma}(h_i))$, para $i=1,2,3$, as estimativas pontuais do semivariograma, obtidas através de um estimador consistente. Se h_1,h_2 e h_3 forem quaisquer três números reais positivos, tais que $h_1<h_2<\phi_0\leq h_3$, então o problema de mínimos quadrados (OLS) tem uma única solução na família dos modelos de tenda.

Demonstração: Seja $\gamma(h;\tau^2,\sigma^2,\phi)$ um semivariograma pertencente à família dos semivariogramas de modelo de tenda, explicitada pela expressão (1.3.5). Devido à relação entre semivariograma e covariograma, $\gamma_0(h)$ e $\gamma(h;\tau^2,\sigma^2,\phi)$ têm função covariograma

da forma

$$C(h;\sigma^2,\phi) = \begin{cases} \sigma^2 - \frac{\sigma^2 h}{\phi} & se \quad 0 \leq h < \phi \\ 0 & se \quad h \geq \phi \end{cases} .$$

Pretende-se determinar o conjunto de soluções do sistema (6.4.2) quando $C(h;\sigma^2,\phi)$ é da forma anterior. Uma vez que a função covariograma é definida por ramos, para explicitar a forma de $C(h;\sigma^2,\phi)$ no sistema, é necessário considerar diferentes casos, consoante a localização de ϕ relativamente a h_1 e h_2.

Quando $\phi \leq h_1$ ou quando $h_1 < \phi \leq h_2$, o sistema resulta num absurdo – de facto, sempre que $\phi \leq h_2$, o covariograma $C(h_2;\sigma^2,\phi)$ é nulo, pela definição de amplitude. Assim, a segunda equação do sistema (6.4.2), resulta em

$$C(h_2;\sigma^2,\phi) = C(h_2;\sigma_0^2,\phi_0) \Leftrightarrow 0 = C(h_2;\sigma_0^2,\phi_0).$$

Como, no modelo de tenda, o covariograma é nulo se e só se é calculado para valores superiores à amplitude, a última equação é verdadeira se e só se $h_2 \geq \phi_0$. Mas, $h_2 \geq \phi_0$ é um absurdo, porque contradiz as hipóteses do teorema. Portanto, o sistema (6.4.2) só pode ter solução, quando $\phi > h_2$. Nesse caso, o sistema (6.4.2) resulta em

$$\begin{cases} C(h_1;\sigma^2,\phi) = C(h_1;\sigma_0^2,\phi_0) \\ C(h_2;\sigma^2,\phi) = C(h_2;\sigma_0^2,\phi_0) \end{cases} \Leftrightarrow \begin{cases} \sigma^2 - \frac{\sigma^2 h_1}{\phi} = \sigma_0^2 - \frac{\sigma_0^2 h_1}{\phi_0} \\ \sigma^2 - \frac{\sigma^2 h_2}{\phi} = \sigma_0^2 - \frac{\sigma_0^2 h_2}{\phi_0} \end{cases} .$$

Trata-se de um sistema com duas equações e duas incógnitas, cuja solução, única, é

$$\begin{cases} \sigma^2 = \sigma_0^2 \\ \phi = \phi_0 \end{cases} .$$

Portanto, para semivariogramas pertencentes ao modelo de tenda, o sistema (6.4.2) é possível e determinado. Assim, uma vez que as condições do **Teorema 6.4.5** estão satisfeitas, o método de mínimos quadrados usuais tem solução única.

□

Corolário 6.4.7. Seja $Z(\mathbf{s})$ um processo com semivariograma isotrópico $\gamma_0(h) = \gamma(h;\tau_0^2,\sigma_0^2,\phi_0)$ definido pelo modelo esférico, e sejam $(h_i,\hat{\gamma}(h_i))$, para $i = 1,2,3$, as estimativas pontuais do semivariograma, obtidas através de um estimador consistente. Se h_1, h_2 e h_3 forem quaisquer três números reais positivos, tais que $h_1 < h_2 < \phi_0 \leq h_3$, então o problema de mínimos quadrados (OLS) tem uma única solução na família dos modelos esféricos.

Demonstração: Seja $\gamma(h;\tau^2,\sigma^2,\phi)$ um semivariograma pertencente à família dos semivariogramas de modelo esférico, explicitada pela expressão (1.3.7). Devido à relação entre semivariograma e covariograma, $\gamma_0(h)$ e $\gamma(h;\tau^2,\sigma^2,\phi)$ têm função covariograma da forma

$$C(h;\sigma^2,\phi)=\begin{cases}\sigma^2\left(1-\frac{3h}{2\phi}+\frac{h^3}{2\phi^3}\right) & se \quad 0\leq h<\phi\\ 0 & se \quad h\geq\phi\end{cases}.$$

Pretende-se determinar o conjunto de soluções do sistema (6.4.2) quando $C(h;\sigma^2,\phi)$ é da forma anterior. Deste modo, se $\phi\leq h_2$, então (6.4.2) resulta num absurdo, pelas razões explicadas na demonstração do corolário anterior.

Considere-se então que $\phi>h_2$. Resolvendo (6.4.2) em ordem às incógnitas σ^2 e ϕ, obtém-se que

$$\begin{cases}C(h_1;\sigma^2,\phi)=C(h_1;\sigma_0^2,\phi_0)\\ C(h_2;\sigma^2,\phi)=C(h_2;\sigma_0^2,\phi_0)\end{cases}\Leftrightarrow\begin{cases}\sigma^2\left(1-\frac{3h_1}{2\phi}+\frac{h_1^3}{2\phi^3}\right)=\sigma_0^2\left(1-\frac{3h_1}{2\phi_0}+\frac{h_1^3}{2\phi_0^3}\right)\\ \sigma^2\left(1-\frac{3h_2}{2\phi}+\frac{h_2^3}{2\phi^3}\right)=\sigma_0^2\left(1-\frac{3h_2}{2\phi_0}+\frac{h_2^3}{2\phi_0^3}\right)\end{cases}.$$

O sistema anterior é equivalente a

$$\begin{cases}\sigma^2=\sigma_0^2\frac{1-\frac{3h_1}{2\phi_0}+\frac{h_1^3}{2\phi_0^3}}{1-\frac{3h_1}{2\phi}+\frac{h_1^3}{2\phi^3}}\\ \frac{2\phi^3-3h_2\phi^2+h_2^3}{2\phi^3-3h_1\phi^2+h_1^3}=\frac{2\phi_0^3-3h_2\phi_0^2+h_2^3}{2\phi_0^3-3h_1\phi_0^2+h_1^3}\end{cases},$$

no entanto, a verificação da unicidade de solução não é imediata. Para resolver a segunda equação do sistema, é necessário confirmar a injectividade da função que define cada um dos membros da equação. Considere-se a função auxiliar

$$\begin{array}{rccc} f\colon & D_f=]h_2,+\infty[& \longrightarrow & \mathbb{R}\\ & \phi & \longmapsto & f(\phi)=\frac{2\phi^3-3h_2\phi^2+h_2^3}{2\phi^3-3h_1\phi^2+h_1^3}\end{array}.$$

Repare-se que o domínio de f não é o conjunto dos números reais uma vez que, como se viu, $\phi>h_2$. Se f for uma função injectiva, então $f(\phi)=f(\phi_0)\Leftrightarrow\phi=\phi_0$ e (6.4.2) fica resolvido.

O polinómio que consta no denominador de f tem zeros em $\phi=h_1$ e em $\phi=-\frac{h_1}{2}$. Como, por hipótese, $0<h_1<h_2$, então esses zeros não pertencem ao domínio da função, isto é, o polinómio no denominador não se anula em D_f. Assim, a função f é contínua e diferenciável, pois é o quociente de dois polinómios cujo denominador não se anula no domínio.

A demonstração prossegue verificando que f é uma função monótona. Para tal, confirmar-se-á que a derivada de f não se anula em D_f.

A derivada de f é a função

$$f'(\phi) = \frac{df(\phi)}{d\phi} = \frac{6\phi\left((\phi - h_2)(2\phi^3 - 3h_1\phi^2 + h_1^3) - (\phi - h_1)(2\phi^3 - 3h_2\phi^2 + h_2^3)\right)}{(2\phi^3 - 3h_1\phi^2 + h_1^3)^2},$$

cujos zeros são dados por

$$\begin{aligned} f'(\phi) = 0 &\Leftrightarrow 6\phi\left((\phi - h_2)(2\phi^3 - 3h_1\phi^2 + h_1^3) - (\phi - h_1)(2\phi^3 - 3h_2\phi^2 + h_2^3)\right) = 0 \\ &\Leftrightarrow \phi = 0 \vee (h_2 - h_1)\phi^3 + (h_1^3 - h_2^3)\phi + h_2^3 h_1 - h_2 h_1^3 = 0 \\ &\Leftrightarrow \phi = 0 \vee \phi = h_1 \vee \phi = h_2 \vee \phi = -\tfrac{h_2^2 - h_1^2}{h_2 - h_1}. \end{aligned}$$

Como nenhum dos zeros da função derivada pertence ao domínio de f, uma vez que $-\frac{h_2^2 - h_1^2}{h_2 - h_1}$, 0 e h_1 são inferiores a h_2 e, por definição, h_2 não pertence a D_f, pode concluir-se que f é estritamente monótona. Assim, ela é uma função injectiva e, por isso,

$$\begin{cases} \sigma^2 = \sigma_0^2 \dfrac{1 - \frac{3h_1}{2\phi_0} + \frac{h_1^3}{2\phi_0^3}}{1 - \frac{3h_1}{2\phi} + \frac{h_1^3}{2\phi^3}} \\ \dfrac{2\phi^3 - 3h_2\phi^2 + h_2^3}{2\phi^3 - 3h_1\phi^2 + h_1^3} = \dfrac{2\phi_0^3 - 3h_2\phi_0^2 + h_2^3}{2\phi_0^3 - 3h_1\phi_0^2 + h_1^3} \end{cases} \Leftrightarrow \begin{cases} \sigma^2 = \sigma_0^2 \\ \phi = \phi_0 \end{cases}.$$

Portanto, para semivariogramas pertencentes ao modelo esférico, o sistema (6.4.2) é possível e determinado. Assim, uma vez que as condições do **Teorema 6.4.5** estão satisfeitas, o método de mínimos quadrados usuais tem solução única.

□

O corolário seguinte apresenta a particularização do **Teorema 6.4.5** para o modelo circular. A demonstração segue a mesma orientação da demonstração do corolário anterior; no entanto, é tecnicamente mais trabalhosa porque a prova da injectividade da função auxiliar não é trivial.

Corolário 6.4.8. Seja $Z(\mathbf{s})$ um processo com semivariograma isotrópico $\gamma_0(h) = \gamma(h; \tau_0^2, \sigma_0^2, \phi_0)$ definido pelo modelo circular, e sejam $(h_i, \hat{\gamma}(h_i))$, para $i = 1, 2, 3$, as estimativas pontuais do semivariograma, obtidas através de um estimador consistente. Se h_1, h_2 e h_3 forem quaisquer três números reais positivos, tais que $h_1 < h_2 < \phi_0 \leq h_3$, então o problema de mínimos quadrados (OLS) tem uma única solução na família dos modelos circulares.

Demonstração: Seja $\gamma(h;\tau^2,\sigma^2,\phi)$ um semivariograma pertencente à família dos semivariogramas de modelo circular, explicitada pela expressão (1.3.6). Devido à relação entre semivariograma e covariograma, $\gamma_0(h)$ e $\gamma(h;\tau^2,\sigma^2,\phi)$ têm função covariograma da forma

$$C(h;\sigma^2,\phi)=\begin{cases}\frac{2\sigma^2}{\pi}\left(\arccos\left(\frac{h}{\phi}\right)-\frac{h}{\phi}\sqrt{1-\frac{h^2}{\phi^2}}\right) & se \quad 0\leq h<\phi\\ 0 & se \quad h\geq\phi\end{cases}.$$

Pretende-se determinar o conjunto de soluções do sistema (6.4.2) quando $C(h;\sigma^2,\phi)$ é da forma anterior. Deste modo, se $\phi\leq h_2$, então (6.4.2) resulta num absurdo, pelas mesmas razões explicadas no **Corolário 6.4.6**.

Considere-se então que $\phi>h_2$. A resolução do sistema (6.4.2), neste caso, corresponde a resolver o sistema

$$\begin{cases}\frac{2\sigma^2}{\pi}\left(\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}\right)=\frac{2\sigma_0^2}{\pi}\left(\arccos\left(\frac{h_1}{\phi_0}\right)-\frac{h_1}{\phi_0}\sqrt{1-\frac{h_1^2}{\phi_0^2}}\right)\\ \frac{2\sigma^2}{\pi}\left(\arccos\left(\frac{h_2}{\phi}\right)-\frac{h_2}{\phi}\sqrt{1-\frac{h_2^2}{\phi^2}}\right)=\frac{2\sigma_0^2}{\pi}\left(\arccos\left(\frac{h_2}{\phi_0}\right)-\frac{h_2}{\phi_0}\sqrt{1-\frac{h_2^2}{\phi_0^2}}\right)\end{cases}$$

em ordem às incógnitas σ^2 e ϕ. O último sistema, pode reescrever-se na forma

$$\begin{cases}\sigma^2=\sigma_0^2\frac{\arccos\left(\frac{h_1}{\phi_0}\right)-\frac{h_1}{\phi_0}\sqrt{1-\frac{h_1^2}{\phi_0^2}}}{\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}}\\ \frac{\arccos\left(\frac{h_2}{\phi}\right)-\frac{h_2}{\phi}\sqrt{1-\frac{h_2^2}{\phi^2}}}{\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}}=\frac{\arccos\left(\frac{h_2}{\phi_0}\right)-\frac{h_2}{\phi_0}\sqrt{1-\frac{h_2^2}{\phi_0^2}}}{\arccos\left(\frac{h_1}{\phi_0}\right)-\frac{h_1}{\phi_0}\sqrt{1-\frac{h_1^2}{\phi_0^2}}}\end{cases}.$$

Passar-se-á a fazer o estudo da função definida pela segunda equação. Assim, seja

$$\begin{array}{rccl} f\colon & D_f=]h_2,+\infty[& \longrightarrow & \mathbb{R} \\ & \phi & \longmapsto & f(\phi)=\frac{\arccos\left(\frac{h_2}{\phi}\right)-\frac{h_2}{\phi}\sqrt{1-\frac{h_2^2}{\phi^2}}}{\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}} \end{array}.$$

O objectivo é verificar que f é uma função injectiva pois, desse modo, $f(\phi)=f(\phi_0)\Leftrightarrow\phi=\phi_0$ e (6.4.2) ficará resolvido.

Tal como na demonstração do corolário anterior, note-se que $D_f\neq\mathbb{R}_0^+$.

Para confirmar a continuidade e a diferenciabilidade de f, verificar-se-á que o denominador de f não tem zeros em D_f.

Para simplificar os cálculos, é conveniente fazer a mudança de variável

$$\begin{array}{rccl} \psi_1\colon &]h_2,+\infty[& \longrightarrow &]0,\frac{\pi}{2}[\\ & \phi & \longmapsto & x=\psi_1(\phi)=\arccos(h_1/\phi) \end{array}.$$

A função ψ_1 é bijectiva e tem inversa $\psi_1^{-1}(x)=\frac{h_1}{\cos(x)}$. Assim,

$$\begin{aligned} \arccos\left(\tfrac{h_1}{\phi}\right)-\tfrac{h_1}{\phi}\sqrt{1-\tfrac{h_1^2}{\phi^2}}=0 &\Leftrightarrow x-\cos(x)\sqrt{1-\cos^2(x)}=0 \\ &\Leftrightarrow x-\cos(x)\sqrt{\sin^2(x)}=0 \end{aligned}$$

Uma vez que $0<h_1/\phi<1$, então $x\in]0,\frac{\pi}{2}[$ e, assim, $\sin(x)>0$. Deste modo, a equação anterior resulta em $x-\cos(x)\sin(x)=g_1(x)=0$. Esta equação tem solução $x=0$ e, dado que $g_1'(x)=2\sin^2(x)>0$ em $x\in]0,\frac{\pi}{2}[$, então $x=0$ é uma solução única. No entanto, por definição da transformação ψ_1, x é necessariamente não nulo; logo a equação anterior é impossível e o denominador de f não se anula em $]h_2,+\infty[$. Sendo assim, f é uma função contínua e diferenciável no seu domínio.

A derivada de f é a função

$$f'(\phi)=\frac{2h_2\sqrt{1-\frac{h_2^2}{\phi^2}}\arccos\left(\frac{h_1}{\phi}\right)-2h_1\sqrt{1-\frac{h_1^2}{\phi^2}}\arccos\left(\frac{h_2}{\phi}\right)}{\phi^2\left(\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}\right)^2},$$

a qual se anula quando

$$\frac{\arccos\left(\frac{h_1}{\phi}\right)}{h_1\sqrt{1-\frac{h_1^2}{\phi^2}}}=\frac{\arccos\left(\frac{h_2}{\phi}\right)}{h_2\sqrt{1-\frac{h_2^2}{\phi^2}}}. \tag{6.4.3}$$

Para resolver a equação (6.4.3) é necessário recorrer ao estudo de uma nova função auxiliar. Seja

$$\begin{array}{rccl} g_2\colon &]0,\phi[& \longrightarrow & \mathbb{R} \\ & h & \longmapsto & g_2(h)=\frac{\arccos\left(\frac{h}{\phi}\right)}{h\sqrt{1-\frac{h^2}{\phi^2}}} \end{array}.$$

Se g_2 for injectiva, então (6.4.3) implica que $h_1=h_2$. Como por hipótese $h_1\neq h_2$, a equação (6.4.3) será impossível.

Assim, os passos seguintes da demonstração têm como objectivo principal a verificação de que g_2 é injectiva, de modo a garantir que f' não tem zeros em D_f.

Para tal, considere-se a derivada de g_2,

$$g_2'(h)=\frac{-\frac{h}{\phi}-\frac{1-2\frac{h^2}{\phi^2}}{\sqrt{1-\frac{h^2}{\phi^2}}}\arccos\left(\frac{h}{\phi}\right)}{h^2\left(1-\frac{h^2}{\phi^2}\right)},$$

que se anula se e só se

$$-\frac{h}{\phi}-\frac{1-2\frac{h^2}{\phi^2}}{\sqrt{1-\frac{h^2}{\phi^2}}}\arccos\left(\frac{h}{\phi}\right)=0. \tag{6.4.4}$$

Para resolver a equação anterior, considere-se a mudança de variável

$$\begin{array}{rccl}\psi_2\colon &]0,\phi[& \longrightarrow &]0,\frac{\pi}{2}[\\ & h & \longmapsto & x=\psi_2(h)=\arccos(h/\phi)\end{array}.$$

ψ_2 é uma função bijectiva e tem inversa $\psi_2^{-1}(x)=\phi\cos(x)$. Assim, a equação anterior é equivalente a

$$-\cos(x)-\frac{1-2\cos^2(x)}{\sqrt{1-\cos^2(x)}}x=0\Leftrightarrow-\cos(x)-\frac{\sin^2(x)-\cos^2(x)}{\sqrt{\sin^2(x)}}x=0.$$

Como $x\in]0,\frac{\pi}{2}[$, então $\sin(x)>0$ e

$$\begin{array}{rcl}-\cos(x)-\frac{\sin^2(x)-\cos^2(x)}{\sqrt{\sin^2(x)}}x=0 & \Leftrightarrow & -\cos(x)\sin(x)+\cos(2x)x=0 \\ & \Leftrightarrow & 2x\cos(2x)=\sin(2x).\end{array}$$

Admitindo que x é diferente de $\pi/4$ (de modo a que $\cos(2x)\neq 0$), a equação anterior é equivalente a $\tan(2x)=2x$. Note-se que não há perda de generalidade ao supor que $x\neq\pi/4$, uma vez que esse valor não é uma solução da equação.

Por outro lado, a função $\tan(2x)-2x$ é uma função crescente em $]-\frac{\pi}{2},\frac{\pi}{2}[$ que só se anula em $x=0$. Mas, por definição de ψ_2, x é não nulo, pelo que a equação anterior é impossível. Consequentemente, a equação (6.4.4) também é impossível, o que vai garantir que a equação (6.4.3) também o seja, isto é, que f' não tem zeros.

Deste modo, mostrou-se que f é injectiva e que o sistema (6.4.2) resulta em

$$\left\{\begin{array}{l}\sigma^2=\sigma_0^2\dfrac{\arccos\left(\frac{h_1}{\phi_0}\right)-\frac{h_1}{\phi_0}\sqrt{1-\frac{h_1^2}{\phi_0^2}}}{\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}} \\ \dfrac{\arccos\left(\frac{h_2}{\phi}\right)-\frac{h_2}{\phi}\sqrt{1-\frac{h_2^2}{\phi^2}}}{\arccos\left(\frac{h_1}{\phi}\right)-\frac{h_1}{\phi}\sqrt{1-\frac{h_1^2}{\phi^2}}}=\dfrac{\arccos\left(\frac{h_2}{\phi_0}\right)-\frac{h_2}{\phi_0}\sqrt{1-\frac{h_2^2}{\phi_0^2}}}{\arccos\left(\frac{h_1}{\phi_0}\right)-\frac{h_1}{\phi_0}\sqrt{1-\frac{h_1^2}{\phi_0^2}}}\end{array}\right.\Leftrightarrow\left\{\begin{array}{l}\sigma=\sigma_0\\ \phi=\phi_0\end{array}\right..$$

Portanto, para semivariogramas pertencentes ao modelo circular, o sistema (6.4.2) é possível e determinado. As condições do **Teorema 6.4.5** estão satisfeitas e o método de mínimos quadrados usuais tem solução única.

□

Note-se que, nos resultados anteriores, foi sempre suposto que $\phi_0 \leq h_3$, o que se traduz pela existência de uma estimativa pontual do semivariograma, determinada num valor de $\|\mathbf{h}\|$ superior ou igual ao valor da amplitude. Esta condição pode ser relaxada para alguns modelos específicos de semivariograma, de acordo com os quais, o método dos mínimos quadrados tem unicidade de solução em condições menos restritivas, nomeadamente, quando todas as estimativas pontuais têm abcissas inferiores à amplitude. Contudo, neste trabalho não se vai seguir essa abordagem.

Concluindo: para qualquer modelo com amplitude, é de considerar duas estimativas pontuais do semivariograma com $\|\mathbf{h}\|$ inferior à amplitude do processo, e outra com $\|\mathbf{h}\|$ igual ou superior.

Os resultados anteriores foram demonstrados apenas para semivariogramas com amplitude, ou seja, apenas para semivariogramas que se tornam funções constantes, iguais ao patamar, a partir de um determinado vector $\mathbf{h}$ de norma finita. No entanto, há um conjunto de semivariogramas pertencentes a processos estacionários de segunda ordem, que só atingem o patamar assintoticamente. Os exemplos mais conhecidos deste tipo de semivariogramas são o modelo exponencial e o modelo Gaussiano, os quais foram apresentados, respectivamente, em (1.3.8) e em (1.3.9). Estes modelos não têm amplitude, mas têm amplitude prática, tal como se referiu na subsecção **1.3.1**. Relembra-se que a amplitude prática é a distância a partir da qual as observações do processo podem ser consideradas não correlacionadas; a partir da amplitude prática, o semivariograma passa a assumir valores muito próximos do patamar. No entanto, os teoremas anteriores não podem ser aplicados directamente a estes casos.

Quando um semivariograma isotrópico tem apenas amplitude prática, é possível que exista um e um só semivariograma que passe por três pontos específicos. Nesse caso, esse semivariograma coincide com o verdadeiro $\gamma_0(h)$. Formalizando a questão nos moldes do estudo que se vem a efectuar, esta situação corresponde aos casos em que

o sistema $\gamma(h_i;\tau^2,\sigma^2,\phi)=\gamma_0(h_i)$, para $i=1,2,3$, é um sistema possível e determinado nas variáveis τ^2,σ^2 e ϕ, para quaisquer $h_1,h_2,h_3\in\mathbb{R}^+$. O teorema seguinte prova que tal acontece no modelo exponencial.

Teorema 6.4.9. Seja $Z(\mathbf{s})$ um processo geoestatístico com semivariograma isotrópico $\gamma_0(h)=\gamma(h;\tau_0^2,\sigma_0^2,\phi_0)$, pertencente à família $\mathfrak{F}$ definida pelo modelo exponencial em (1.3.8), e sejam h_1,h_2 e h_3 quaisquer três números reais positivos distintos. Então, existe apenas um semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_1,\gamma_0(h_1)),(h_2,\gamma_0(h_2))$ e $(h_3,\gamma_0(h_3))$, que é o próprio $\gamma_0(h)$.

Demonstração: Sejam $\gamma_0(h)$ e $\gamma(h;\tau^2,\sigma^2,\phi)$ elementos da família $\mathfrak{F}$, e $h_1,h_2,h_3\in\mathbb{R}^+$ tais que $h_1<h_2<h_3$. O semivariograma $\gamma(h;\tau^2,\sigma^2,\phi)$ passa pelos pontos $(h_i,\gamma_0(h_i))$, $i=1,2,3$, se e só se $\gamma(h_i;\tau^2,\sigma^2,\phi)=\gamma_0(h_i)$, para $i=1,2,3$.

Atendendo à forma de $\gamma(h;\tau^2,\sigma^2,\phi)$, os parâmetros τ^3,σ^2 e ϕ são as incógnitas do sistema

$$\begin{cases}\tau^2+\sigma^2\left(1-e^{-3\frac{h_1}{\phi}}\right)=\tau_0^2+\sigma_0^2\left[1-e^{-3\frac{h_1}{\phi_0}}\right]\\ \tau^2+\sigma^2\left(1-e^{-3\frac{h_2}{\phi}}\right)=\tau_0^2+\sigma_0^2\left[1-e^{-3\frac{h_2}{\phi_0}}\right]\\ \tau^2+\sigma^2\left(1-e^{-3\frac{h_3}{\phi}}\right)=\tau_0^2+\sigma_0^2\left[1-e^{-3\frac{h_3}{\phi_0}}\right]\end{cases}.$$

Subtraindo, membro a membro, a primeira equação, à segunda equação e à terceira equação, respectivamente, o sistema anterior reescreve-se na forma

$$\begin{cases}\tau^2+\sigma^2\left(1-e^{-3\frac{h_1}{\phi}}\right)=\tau_0^2+\sigma_0^2\left[1-e^{-3\frac{h_1}{\phi_0}}\right]\\ \sigma^2\left(e^{-3\frac{h_2}{\phi}}-e^{-3\frac{h_1}{\phi}}\right)=\sigma_0^2\left(e^{-3\frac{h_2}{\phi_0}}-e^{-3\frac{h_1}{\phi_0}}\right)\\ \sigma^2\left(e^{-3\frac{h_3}{\phi}}-e^{-3\frac{h_1}{\phi}}\right)=\sigma_0^2\left(e^{-3\frac{h_3}{\phi_0}}-e^{-3\frac{h_1}{\phi_0}}\right)\end{cases}.$$

Por substituição, obtém-se que

$$\begin{cases}\tau^2=\tau_0^2+\sigma_0^2\left[1-e^{-3\frac{h_1}{\phi_0}}\right]-\sigma^2\left(1-e^{-3\frac{h_1}{\phi}}\right)\\ \sigma^2=\sigma_0^2\dfrac{e^{-3\frac{h_2}{\phi_0}}-e^{-3\frac{h_1}{\phi_0}}}{e^{-3\frac{h_2}{\phi}}-e^{-3\frac{h_1}{\phi}}}\\ \dfrac{e^{-3\frac{h_3}{\phi}}-e^{-3\frac{h_1}{\phi}}}{e^{-3\frac{h_2}{\phi}}-e^{-3\frac{h_1}{\phi}}}=\dfrac{e^{-3\frac{h_3}{\phi_0}}-e^{-3\frac{h_1}{\phi_0}}}{e^{-3\frac{h_2}{\phi_0}}-e^{-3\frac{h_1}{\phi_0}}}\end{cases},$$

que é equivalente a

$$\begin{cases} \tau^2 = \tau_0^2 + \sigma_0^2\left[1 - e^{-3\frac{h_1}{\phi_0}}\right] - \sigma^2\left(1 - e^{-3\frac{h_1}{\phi}}\right) \\ \sigma^2 = \sigma_0^2 \frac{e^{-3\frac{h_2}{\phi_0}} - e^{-3\frac{h_1}{\phi_0}}}{e^{-3\frac{h_2}{\phi}} - e^{-3\frac{h_1}{\phi}}} \\ \frac{\left(e^{\frac{-3}{\phi}}\right)^{h_3-h_1} - 1}{\left(e^{\frac{-3}{\phi}}\right)^{h_2-h_1} - 1} = \frac{\left(e^{\frac{-3}{\phi_0}}\right)^{h_3-h_1} - 1}{\left(e^{\frac{-3}{\phi_0}}\right)^{h_2-h_1} - 1} \end{cases}.$$

Assim, utilizando técnicas idênticas às das demonstrações dos corolários **6.4.7** e **6.4.8**, passa-se a apresentar o estudo das soluções da terceira equação do sistema.

Efectuando a mudança de variável

$$\begin{array}{rccl} \psi\colon &]0,+\infty[& \longrightarrow &]0,1[\\ & \phi & \longmapsto & x = \psi(\phi) = e^{\frac{-3}{\phi}} \end{array},$$

verifica-se que a terceira equação do sistema anterior resulta em

$$\frac{x^{h_3-h_1} - 1}{x^{h_2-h_1} - 1} = \frac{x_0^{h_3-h_1} - 1}{x_0^{h_2-h_1} - 1}, \tag{6.4.5}$$

pois ψ é uma função bijectiva e tem inversa $\psi^{-1}(x) = \frac{-3}{\ln(x)}$.

Para resolver (6.4.5), considere-se a função

$$\begin{array}{rccl} f\colon & D_f =]0,1[& \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) = \frac{x^{h_3-h_1-1}}{x^{h_2-h_1-1}} \end{array}.$$

Se f for injectiva, a equação (6.4.5) fica automaticamente resolvida, o que se passará a comprovar.

O denominador de f é um polinómio que não se anula em D_f porque, como $h_2 > h_1$, então $0 < x^{h_2-h_1} < 1$. Sendo assim, f é uma função contínua e diferenciável, com derivada

$$f'(x) = \frac{(h_3-h_1)x^{h_3-h_1-1}(x^{h_2-h_1}-1) - (h_2-h_1)x^{h_2-h_1-1}(x^{h_3-h_1}-1)}{(x^{h_2-h_1}-1)^2}.$$

Por reorganização dos termos no numerador, é possível verificar que

$$f'(x) = 0 \Leftrightarrow (h_3-h_2)x^{h_3-h_1} - (h_3-h_1)x^{h_3-h_2} + h_2 - h_1 = 0. \tag{6.4.6}$$

Seja $g(x) = (h_3-h_2)x^{h_3-h_1} - (h_3-h_1)x^{h_3-h_2} + h_2 - h_1$, quando $x \in D_f$. Ora, $g(0) = h_2 - h_1$, que, por hipótese, é um número positivo, e $g(1) = 0$. Interessa verificar que g é estritamente decrescente em D_f.

A função g tem derivada

$$g'(x) = (h_3 - h_2)(h_3 - h_1)x^{h_3-h_1-1} - (h_3 - h_1)(h_3 - h_2)x^{h_3-h_2-1},$$

a qual se anula se e só se

$$x^{h_3-h_2-1}\left(x^{h_2-h_1} - 1\right) = 0.$$

Como $x \in]0, 1[$, então a equação anterior é impossível e g' não tem zeros em D_f, ou seja, g é uma função estritamente decrescente em D_f.

Por outro lado, como $g(1) = 0$, para todo o $x \in]0, 1[$, pode-se concluir que $g(x) \neq 0$. Logo, a equação (6.4.6) é impossível e, por isso, f' não se anula em D_f. Assim f é uma função injectiva e a equação (6.4.5) tem como única solução $x = x_0 \Leftrightarrow e^{\frac{-3}{\phi}} = e^{\frac{-3}{\phi_0}}$.

Como a função exponencial é injectiva, $e^{\frac{-3}{\phi}} = e^{\frac{-3}{\phi_0}} \Leftrightarrow \phi = \phi_0$ e o sistema inicial resulta em

$$\begin{cases} \gamma(h_1; \tau^2, \sigma^2, \phi) = \gamma_0(h_1) \\ \gamma(h_2; \tau^2, \sigma^2, \phi) = \gamma_0(h_2) \\ \gamma(h_3; \tau^2, \sigma^2, \phi) = \gamma_0(h_3) \end{cases} \Leftrightarrow \begin{cases} \tau^2 = \tau_0^2 \\ \sigma^2 = \sigma_0^2 \\ \phi = \phi_0 \end{cases}.$$

Portanto, o único semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_i, \gamma_0(h_i))$, para $i = 1, 2, 3$, é o próprio $\gamma_0(h)$.

□

Corolário 6.4.10. Sejam $(h_i, \hat{\gamma}(h_i))$, para $i = 1, 2, 3$, as estimativas pontuais do semivariograma, obtidas a partir de um estimador consistente. Nas condições do **Teorema 6.4.9**, e para uma dimensão amostral suficientemente grande, o problema de mínimos quadrados (OLS) tem uma única solução, com modelo de semivariograma exponencial.

Demonstração: Decorre do **Teorema 6.4.9**, utilizando uma demonstração análoga à do **Teorema 6.4.5**.

□

Os resultados anteriores garantem que, teoricamente, o modelo exponencial tem uma única solução de mínimos quadrados, quando se utilizam quaisquer três estimativas pontuais do semivariograma para estimar os parâmetros do modelo. No entanto, na prática, existem problemas numéricos que fazem com que não se possa utilizar esta propriedade. De facto, como a partir da amplitude prática, o semivariograma é

muito próximo do patamar, se h for significativamente superior à amplitude prática verifica-se, frequentemente, que os métodos computacionais/numéricos não são capazes de distinguir entre a aproximação ao patamar $\gamma_0(h)$ e o próprio patamar $\tau_0^2 + \sigma_0^2$. Por isso, mesmo que, teoricamente, exista apenas um semivariograma $\gamma_0(h)$ que passa por três pontos específicos, em termos práticos, a situação é semelhante à verificada nos modelos de semivariograma com amplitude. Portanto, se existir no máximo um dos h_i, com $i = 1, 2, 3$, menor do que a amplitude prática de $\gamma_0(h)$, é provável que o método dos mínimos quadrados usuais também devolva soluções múltiplas.

A Figura 6.7 exemplifica a situação que se acabou de referir. Nela pode-se visualizar um conjunto de semivariogramas distintos de modelo exponencial que, na prática, são todos soluções de mínimos quadrados usuais. Utilizando o *software R* com a *package geoR*, todos os semivariogramas da figura passam pelas três estimativas pontuais do semivariograma consideradas. Este conjunto de soluções ocorre, pelo facto de existir apenas uma estimativa pontual com abcissa inferior à amplitude prática do semivariograma verdadeiro de $Z(\mathbf{s})$. Além disso, as outras duas estimativas pontuais estão muito próximas do patamar, o que leva o *software* a considerá-las iguais ao patamar.

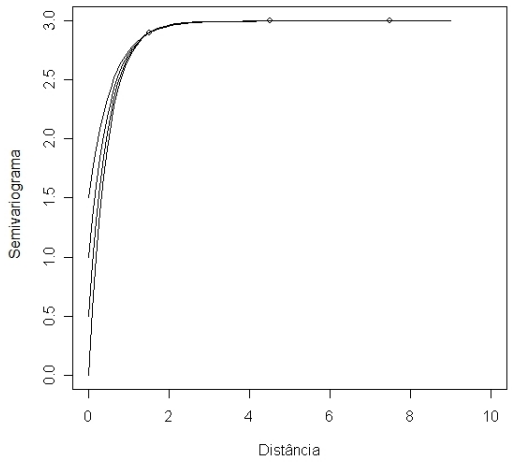

Figura 6.7: Ilustração da existência de soluções múltiplas, na estimação dos parâmetros de um modelo de semivariograma exponencial, pelo método dos mínimos quadrados usuais.

Do conjunto de resultados até agora demonstrados, pode-se concluir que, se $Z(\mathbf{s})$ for um processo estacionário de segunda ordem, em termos práticos devem ser sempre consideradas duas estimativas pontuais com norma de $\mathbf{h}$ inferior à amplitude do processo $Z(\mathbf{s})$ (ou à amplitude prática, se for o caso) e outra com norma de $\mathbf{h}$ superior. Procedendo dessa forma, evitam-se as soluções múltiplas no método de mínimos quadrados.

Portanto, é necessário conhecer inicialmente uma boa aproximação da amplitude do processo (ou da amplitude prática), para que se possa estimar pontualmente o semivariograma nos vectores adequados. A estimativa inicial da amplitude (ou da amplitude prática), deve ser obtida por um método robusto.

Apesar dos resultados obtidos até agora se revelarem bastante úteis em processos estacionários de segunda ordem, eles não se podem aplicar em processos apenas intrinsecamente estacionários. De facto, os processos que são apenas intrinsecamente estacionários, nem sequer têm um parâmetro análogo à amplitude (ou à amplitude prática) e, por isso, os resultados obtidos deixam de fazer qualquer sentido nesses modelos. Contudo, os processos que são apenas intrinsecamente estacionários têm a vantagem de possuir semivariogramas ilimitados. Essa propriedade facilita a existência de solução única.

Seguidamente, considerar-se-á o modelo de potência, que é um dos modelos mais utilizados em processos apenas intrinsecamente estacionários.

Teorema 6.4.11. Seja $Z(\mathbf{s})$ um processo geoestatístico com semivariograma isotrópico $\gamma_0(h) = \gamma(h; \tau_0^2, \theta_0, \lambda_0)$, pertencente à família $\mathfrak{F}$ definida pelo modelo de potência em (1.3.10), e sejam h_1, h_2 e h_3 quaisquer três números reais positivos distintos. Então, existe apenas um semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_1, \gamma_0(h_1)), (h_2, \gamma_0(h_2))$ e $(h_3, \gamma_0(h_3))$, que é o próprio $\gamma_0(h)$.

Demonstração: Sejam $\gamma_0(h)$ e $\gamma(h; \tau^2, \theta, \lambda)$ elementos da família $\mathfrak{F}$, e $h_1, h_2, h_3 \in \mathbb{R}^+$ tais que $h_1 < h_2 < h_3$. O semivariograma $\gamma(h; \tau^2, \theta, \lambda)$ passa pelos pontos $(h_i, \gamma_0(h_i))$, $i = 1, 2, 3$, se e só se $\gamma(h_i; \tau^2, \theta, \lambda) = \gamma_0(h_i)$, para $i = 1, 2, 3$.

Atendendo à forma de $\gamma(h; \tau^2, \theta, \lambda)$, os parâmetros τ^2, θ e λ são as incógnitas do

sistema

$$\begin{cases} \tau^2 + \theta h_1^{\lambda} = \tau_0^2 + \theta_0 h_1^{\lambda_0} \\ \tau^2 + \theta h_2^{\lambda} = \tau_0^2 + \theta_0 h_2^{\lambda_0} \\ \tau^2 + \theta h_3^{\lambda} = \tau_0^2 + \theta_0 h_3^{\lambda_0} \end{cases} .$$

Subtraindo, membro a membro, a primeira equação, à segunda equação e à terceira equação, respectivamente, o sistema anterior reescreve-se na forma

$$\begin{cases} \tau^2 = \tau_0^2 + \theta_0 h_1^{\lambda_0} - \theta h_1^{\lambda} \\ \theta \left(h_2^{\lambda} - h_1^{\lambda}\right) = \theta_0 \left(h_2^{\lambda_0} - h_1^{\lambda_0}\right) \\ \theta \left(h_3^{\lambda} - h_1^{\lambda}\right) = \theta_0 \left(h_3^{\lambda_0} - h_1^{\lambda_0}\right) \end{cases} .$$

Por substituição, obtém-se que

$$\begin{cases} \tau^2 = \tau_0^2 + \theta_0 h_1^{\lambda_0} - \theta h_1^{\lambda} \\ \theta = \theta_0 \frac{h_2^{\lambda_0} - h_1^{\lambda_0}}{h_2^{\lambda} - h_1^{\lambda}} \\ \frac{h_3^{\lambda} - h_1^{\lambda}}{h_2^{\lambda} - h_1^{\lambda}} = \frac{h_3^{\lambda_0} - h_1^{\lambda_0}}{h_2^{\lambda_0} - h_1^{\lambda_0}} \end{cases} ,$$

o que é equivalente ao sistema

$$\begin{cases} \tau^2 = \tau_0^2 + \theta_0 h_1^{\lambda_0} - \theta h_1^{\lambda} \\ \theta = \theta_0 \frac{h_2^{\lambda_0} - h_1^{\lambda_0}}{h_2^{\lambda} - h_1^{\lambda}} \\ \frac{\left(e^{\lambda}\right)^{\ln h_3} - \left(e^{\lambda}\right)^{\ln h_1}}{\left(e^{\lambda}\right)^{\ln h_2} - \left(e^{\lambda}\right)^{\ln h_1}} = \frac{\left(e^{\lambda_0}\right)^{\ln h_3} - \left(e^{\lambda_0}\right)^{\ln h_1}}{\left(e^{\lambda_0}\right)^{\ln h_2} - \left(e^{\lambda_0}\right)^{\ln h_1}} \end{cases} .$$

Assim, utilizando técnicas idênticas às do **Teorema 6.4.9**, passa-se a apresentar o estudo das soluções da terceira equação do sistema.

Efectuando a mudança de variável

$$\begin{array}{rccl} \psi\colon &]0,2[& \longrightarrow &]1,e^2[\\ & \lambda & \longmapsto & x = \psi(\lambda) = e^{\lambda} \end{array} ,$$

verifica-se que a terceira equação do sistema anterior resulta em

$$\frac{x^{\ln h_3 - \ln h_1} - 1}{x^{\ln h_2 - \ln h_1} - 1} = \frac{x_0^{\ln h_3 - \ln h_1} - 1}{x_0^{\ln h_2 - \ln h_1} - 1}, \tag{6.4.7}$$

pois ψ é uma função bijectiva e tem inversa $\psi^{-1}(x) = \ln(x)$.

Para resolver (6.4.7), considere-se a função auxiliar

$$\begin{array}{rccl} f\colon & D_f =]1,e^2[& \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x) = \frac{x^{\ln h_3 - \ln h_1} - 1}{x^{\ln h_2 - \ln h_1} - 1} \end{array} .$$

Se f for injectiva, a equação (6.4.7) fica automaticamente resolvida, como se passará a comprovar.

O denominador de f é um polinómio que não se anula em D_f porque, como $h_2 > h_1$, então $1 < x^{\ln h_2 - \ln h_1}$. Sendo assim, f é uma função contínua e diferenciável, com derivada $f'(x) =$

$$\frac{(\ln h_3 - \ln h_1)x^{\ln h_3 - \ln h_1 - 1}(x^{\ln h_2 - \ln h_1} - 1) - (\ln h_2 - \ln h_1)x^{\ln h_2 - \ln h_1 - 1}(x^{\ln h_3 - \ln h_1} - 1)}{(x^{\ln h_2 - \ln h_1} - 1)^2},$$

que se anula quando

$$(\ln h_3 - \ln h_2)x^{\ln h_3 - \ln h_1} - (\ln h_3 - \ln h_1)x^{\ln h_3 - \ln h_2} + \ln h_2 - \ln h_1 = 0. \tag{6.4.8}$$

Para resolver a equação anterior, seja

$$g(x) = (\ln h_3 - \ln h_2)x^{\ln h_3 - \ln h_1} - (\ln h_3 - \ln h_1)x^{\ln h_3 - \ln h_2} + \ln h_2 - \ln h_1$$

quando $x \in D_f$. Ora, $g(1) = 0$ e g é uma função contínua e diferenciável com derivada

$$g'(x) = (\ln h_3 - \ln h_2)(\ln h_3 - \ln h_1)x^{\ln h_3 - \ln h_1 - 1} - (\ln h_3 - \ln h_1)(\ln h_3 - \ln h_2)x^{\ln h_3 - \ln h_2 - 1},$$

a qual se anula se e só se

$$x^{\ln h_3 - \ln h_2 - 1}\left(x^{\ln h_2 - \ln h_1} - 1\right) = 0.$$

Como $x \in]1, e^2[$, então a equação anterior é impossível e g' não tem zeros em D_f.

Assim, g é uma função estritamente monótona em D_f e, como $g(1) = 0$, então, para todo o $x \in]1, e^2[$, $g(x) \neq 0$. A equação (6.4.8) é impossível e, por isso, f' não se anula em D_f. Deste modo, f é uma função injectiva e a equação (6.4.7) tem como única solução $x = x_0 \Leftrightarrow e^{\lambda} = e^{\lambda_0} \Leftrightarrow \lambda = \lambda_0$. Logo, o sistema inicial resulta em

$$\begin{cases} \gamma(h_1; \tau^2, \theta, \lambda) = \gamma_0(h_1) \\ \gamma(h_2; \tau^2, \theta, \lambda) = \gamma_0(h_2) \\ \gamma(h_3; \tau^2, \theta, \lambda) = \gamma_0(h_3) \end{cases} \Leftrightarrow \begin{cases} \tau^2 = \tau_0^2 \\ \theta = \theta_0 \\ \lambda = \lambda_0 \end{cases}.$$

Portanto, o único semivariograma de $\mathfrak{F}$ que passa pelos pontos $(h_i, \gamma_0(h_i))$, para $i = 1, 2, 3$, é o próprio $\gamma_0(h)$.

□

Corolário 6.4.12. Sejam $(h_i, \hat{\gamma}(h_i))$, para $i = 1, 2, 3$, as estimativas pontuais do semivariograma, obtidas através de um estimador consistente. Nas condições do **Teorema 6.4.11**, e para uma dimensão amostral suficientemente grande, o problema de mínimos quadrados (OLS) tem uma única solução, com modelo de semivariograma de potência.

Demonstração: Decorre do **Teorema 6.4.11**, utilizando uma demonstração análoga à do **Teorema 6.4.5**.

□

Os resultados anteriores mostram que, teoricamente, o modelo de potência tem uma única solução de mínimos quadrados, quando se utilizam quaisquer três estimativas pontuais do semivariograma, independentemente da sua localização em relação aos parâmetros do modelo. Os semivariogramas do modelo de potência não têm patamar e não são limitados superiormente. Consequentemente, os problemas computacionais que se verificaram, por exemplo, no modelo exponencial, não existem no modelo de potência. Portanto, quer teoricamente, quer na prática, o método dos mínimos quadrados usuais tem apenas uma solução na família dos modelos de potência.

O resultado do estudo efectuado sobre as condições que garantem a identificabilidade dos parâmetros do modelo de semivariograma e, consequentemente, a unicidade de solução pelo método dos mínimos quadrados, será incorporado no processo de estimação dos múltiplos variogramas, a descrever sumariamente no algoritmo que se apresenta na subsecção seguinte.

6.4.3 Algoritmo para o cálculo de estimativas

Considere-se um processo estacionário de segunda ordem, com variograma isotrópico pertencente a uma das famílias consideradas no presente capítulo. Seja q o número de parâmetros do modelo ($q = 2, 3$). O cálculo das estimativas de acordo com o estimador de múltiplos variogramas resume-se nos passos seguintes.

Algoritmo:

1. Determina-se uma aproximação inicial da amplitude do processo, $\hat{\phi}_{\text{ini}}$, através de um procedimento robusto (por exemplo, usando o estimador Q_n de Genton

associado ao método de OLS).

2. Através do estimador Q_n de Genton, calculam-se $q-1$ estimativas pontuais do variograma com abcissas inferiores a $\hat{\phi}_{\text{ini}}$ e uma estimativa com abcissa não inferior a esse valor.

3. Pelo método dos mínimos quadrados (OLS), estimam-se os parâmetros do modelo $2\gamma(\|\mathbf{h}\|, \boldsymbol{\theta})$, utilizando as estimativas pontuais encontradas no passo anterior. Deste modo, obtém-se o vector de estimativas dos parâmetros $\hat{\boldsymbol{\theta}}_b = (\hat{\theta}_{b,1}, ..., \hat{\theta}_{b,q})$.

4. Variando as abcissas onde se estima pontualmente o variograma, repetem-se os passos 2. e 3., B vezes, para obter B estimativas, $(\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_B)$, dos parâmetros do modelo.

5. A estimativa final do variograma é dada por $2\gamma(\|\mathbf{h}\|, \tilde{\boldsymbol{\theta}})$, onde

$$\tilde{\boldsymbol{\theta}} = \left(\text{Mediana}\{\hat{\theta}_{1,1}, ..., \hat{\theta}_{B,1}\}, ..., \text{Mediana}\{\hat{\theta}_{1,q}, ..., \hat{\theta}_{B,q}\}\right).$$

Se o processo considerado for apenas intrinsecamente estacionário, o ponto 1. do algoritmo não tem sentido, uma vez que não existe amplitude. Consequentemente, nesses casos, o algoritmo reduz-se aos passos seguintes, e as estimativas pontuais podem ser determinadas em qualquer vector $\mathbf{h}$ pretendido.

Por outro lado, quando o processo não tem um semivariograma isotrópico, o procedimento anterior deve ser utilizado em cada uma das direcções onde se deseja estimar o semivariograma.

6.4.4 Propriedades assintóticas

Nesta subsecção estudam-se as propriedades assintóticas do estimador de múltiplos variogramas. Para tal, é necessário considerar, separadamente, cada etapa do método, isto é, é preciso estudar a estimação pontual do variograma (que é feita através do estimador Q_n de Genton), a estimação repetida dos parâmetros do modelo de variograma (que é feita por OLS) e, por fim, a estimação da mediana do conjunto de estimativas dos parâmetros.

Comece-se, então, por apresentar as propriedades assintóticas do estimador Q_n de Genton, o qual foi definido em (6.2.4). Antes de mais, é necessário recordar as propriedades assintóticas do estimador Q_n, inicialmente proposto por Rousseeuw e Croux (1993) no contexto de estimação de escala. A proposição que se apresenta de seguida, decorre de resultados publicados no artigo referido, assumindo as condições de regularidade necessárias. Ainda no mesmo artigo, pode ser encontrada a função de influência do estimador Q_n.

Proposição 6.4.13. Seja $(X_1, ..., X_n)$ uma amostra aleatória de uma população com *f.d.p.* F e função densidade de probabilidade f. Considere-se o estimador $\hat{\sigma} = Q_n(X_1, ..., X_n)$ de $\sigma = \sqrt{\text{Var}[X_i]}$, apresentado em (3.3.6), e duas variáveis aleatórias X e Y, *i.i.d.* com *f.d.p.* F. Seja c a constante que torna o estimador Q_n consistente (*vide* ponto (3.3.6)). Se $f_{X-Y}(1/c) > 0$ então

$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{\mathcal{L}} N\left(0, 4\sigma^2 \text{Var}[Q, F]\right),$$

onde $\mathcal{L}$ representa convergência em lei, $\text{Var}[Q, F] = \int IF(x; Q, F)^2 dF(x)$ e $IF(x; Q, F)$ é a função de influência do estimador Q_n.

Demonstração: Com base em resultados de Serfling (1984), Rousseeuw e Croux concluíram que o estimador $\hat{\sigma}$ converge em lei para uma distribuição normal, nomeadamente,

$$\sqrt{n}\left(\hat{\sigma} - \sigma\right) \xrightarrow{\mathcal{L}} N\left(0, \text{Var}[Q, F]\right).$$

Para concluir a convergência em lei do estimador da variância, basta aplicar o Método Delta ao desenvolvimento da função $g(\sigma) = \sigma^2$, o qual garante que

$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) \xrightarrow{\mathcal{L}} N\left(0, (g'(\sigma))^2 \text{Var}[Q, F]\right),$$

para funções g com derivada finita diferente de zero.

Como $(g'(\sigma))^2 = 4\sigma^2 > 0$, decorre de imediato o resultado que se pretende demonstrar.

□

O resultado da proposição anterior permite tirar conclusões acerca da consistência do estimador Q_n.

Corolário 6.4.14. Nas condições da **Proposição 6.4.13**, $Q_n^2(X_1, ..., X_n)$ é um estimador consistente de σ^2, isto é,

$$Q_n^2(X_1, ..., X_n) \xrightarrow{\mathcal{P}} \sigma^2,$$

onde $\mathcal{P}$ representa convergência em probabilidade.

Demonstração: Sabe-se que, se $\{k_n\}_{n\in\mathbb{N}}$ é uma sucessão de constantes que tende para infinito e $k_n(T_n - \theta) \xrightarrow{\mathcal{L}} F$, onde T_n é um estimador de θ e F é uma *f.d.p.* contínua, então $T_n \xrightarrow{\mathcal{P}} \theta$.

Tomando $T_n = Q_n^2(X_1, ..., X_n)$, $\theta = \sigma^2$ e $k_n = \sqrt{n}$, conclui-se o resultado pretendido.

□

Os resultados anteriores foram demonstrados num contexto geral, não espacial. Em processos espaciais, é necessário especificar qual é o contexto assintótico em que se trabalha (*IA*, *IDA* ou *MIDA*, de acordo com a subsecção **2.3.1**). Dependendo do contexto, o estimador pode gozar de diferentes propriedades. Para estudar as propriedades de Q_n na primeira fase do estimador de múltiplos variogramas, considerar-se-ão os cenários *IDA* e *MIDA*.

No seguimento será suposta uma hipótese adicional sobre a região inicial R_0 referida na subsecção **2.3.1**, de modo a excluir formas anómalas do domínio do processo. Recorre-se à notação usual de $O(.)$ e de $o(.)$, segundo a qual, $u_n = O(v_n)$ significa que u_n/v_n é uma sucessão limitada, e que $u_n = o(v_n)$ significa que $u_n/v_n \underset{n\to\infty}{\longrightarrow} 0$.

C.1 Para qualquer sequência de números reais positivos $\{t_n\}_{n\in\mathbb{N}}$, com $t_n \underset{n\to\infty}{\longrightarrow} 0$, o número de conjuntos da forma $(\mathbf{i}+]0,1]^d)t_n$ (com $\mathbf{i} = (i_1, ..., i_d) \in \mathbb{Z}^d$) que intersectam R_0 e o seu complementar, é $O(t_n^{d-1})$ quando $n \to \infty$.

A hipótese, que se verifica em todos os processos com interesse prático, também foi assumida por Lahiri *et al.* (2002) e, de acordo com os autores, garante que o número de incrementos separados por um dado vector fixo e o número de observações da amostra, tendem para infinito com a mesma velocidade de convergência. Assim, quando $n \to \infty$, e para qualquer $\mathbf{h} \in \mathfrak{X}^d$,

$$\#N_n(\mathbf{h}) = n(1 + o(1)).$$

Este resultado vai ser utilizado na demonstração do próximo teorema.

Teorema 6.4.15. Seja $Z(\mathbf{s})$ um processo geoestatístico com variograma $2\gamma(\mathbf{h}, \boldsymbol{\theta}_0) = 2\gamma_0(\mathbf{h})$ e $\{Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)\}$ uma sua amostra. Para um qualquer vector $\mathbf{h} \in \mathfrak{X}^d$ fixo, seja $2\hat{\gamma}_G(\mathbf{h})$ o estimador Q_n de Genton apresentado em (6.2.4). Considere-se que os incrementos $I_i = \{Z(\mathbf{s}_i) - Z(\mathbf{s}_i + \mathbf{h}) : i = 1, ..., \#N(\mathbf{h})\}$ são independentes, que a função densidade $f_{I_i - I_j}(1/c) > 0$ e que C.1 se verifica. Então, sobre condições de *IDA* ou de *MIDA* verifica-se que

- $\sqrt{n}\left(2\hat{\gamma}_G(\mathbf{h}) - 2\gamma_0(\mathbf{h})\right) \xrightarrow{\mathcal{L}} N\left(0, 8\gamma_0(\mathbf{h})\mathrm{Var}[Q, F]\right)$;
- $2\hat{\gamma}_G(\mathbf{h}) \xrightarrow{\mathcal{P}} 2\gamma_0(\mathbf{h})$.

Demonstração: Como $f_{I_i - I_j}(1/c) > 0$ e os incrementos são independentes, pode-se aplicar directamente o **Proposição 6.4.13** aos incrementos $\{I_i : i = 1, ..., \#N_n(\mathbf{h})\}$, para concluir que

$$\sqrt{\#N_n(\mathbf{h})}\left(2\hat{\gamma}_G(\mathbf{h}) - 2\gamma_0(\mathbf{h})\right) \xrightarrow{\mathcal{L}} N\left(0, 8\gamma_0(\mathbf{h})\mathrm{Var}[Q, F]\right). \tag{6.4.9}$$

Por outro lado, decorre da hipótese C.1 que $\#N_n(\mathbf{h}) = n(1 + o(1))$, quando $n \to \infty$, em qualquer um dos contextos assintóticos considerados. Isto garante que

$$\lim_{n\to\infty} \frac{\#N_n(\mathbf{h})}{n} = 1,$$

isto é, que $\#N_n(\mathbf{h}) = O(n)$. Tendo em conta este resultado e a expressão (6.4.9) pode-se concluir que

$$\sqrt{n}\left(2\hat{\gamma}_G(\mathbf{h}) - 2\gamma_0(\mathbf{h})\right) \xrightarrow{\mathcal{L}} N\left(0, 8\gamma_0(\mathbf{h})\mathrm{Var}[Q, F]\right).$$

Aplicando agora o **Corolário 6.4.14** obtém-se que

$$2\hat{\gamma}_G(\mathbf{h}) \xrightarrow{\mathcal{P}} 2\gamma_0(\mathbf{h}),$$

e fica assim demonstrado o teorema.

□

Antes de prosseguir, analisam-se as hipóteses colocadas no **Teorema 6.4.15**. A hipótese que garante que $f_{I_i - I_j}(1/c) > 0$ não é muito restritiva. De facto, se o processo $Z(\mathbf{s})$ for Gaussiano, essa hipótese verifica-se sempre. No entanto, a hipótese

dos incrementos serem independentes, é demasiado forte. Realmente, os processos geoestatísticos não verificam esta condição. Apesar disso, este teorema constitui um resultado que se verifica aproximadamente em situações de fraca dependência dos incrementos. Genton (1998b) utilizou uma técnica semelhante a esta, para aproximar a matriz de covariâncias do estimador Q_n, em situações onde existe estrutura de dependência. O autor supôs que as observações do processo são independentes para poder aproximar a situação onde existe dependência. Foi nas mesmas condições que o autor concluiu a consistência do estimador Q_n. Assim, prossegue-se o estudo das propriedades do estimador de múltiplos variogramas, utilizando o **Teorema 6.4.15** como um resultado aproximado, tendo em atenção que seria desejável a demonstração de um resultado idêntico com hipóteses menos restritivas.

Corolário 6.4.16. Seja $(2\hat{\gamma}_G(\mathbf{h}_1), ..., 2\hat{\gamma}_G(\mathbf{h}_H))$ o estimador Q_n de Genton considerado nos vectores $\mathbf{h}_i, i = 1, ..., H$. Se para qualquer $\mathbf{h}_i$, se verificarem as condições do teorema anterior, então

$$(2\hat{\gamma}_G(\mathbf{h}_1), ..., 2\hat{\gamma}_G(\mathbf{h}_H)) \xrightarrow{\mathcal{P}} (2\gamma_0(\mathbf{h}_1), ..., 2\gamma_0(\mathbf{h}_H)).$$

Demonstração: É imediata utilizando a definição de convergência em probabilidade multidimensional.

□

De modo idêntico, a distribuição assintótica normal multivariada do estimador é um corolário do **Teorema 6.4.15**.

Corolário 6.4.17. Seja $(2\hat{\gamma}_G(\mathbf{h}_1), ..., 2\hat{\gamma}_G(\mathbf{h}_H))$ o estimador Q_n de Genton considerado nos vectores $\mathbf{h}_i, i = 1, ..., H$. Se para qualquer $\mathbf{h}_i$, se verificarem as condições do **Teorema 6.4.15**, então

$$\sqrt{n}\left(2\hat{\gamma}_G(\mathbf{h}_1) - 2\gamma(\mathbf{h}_1, \boldsymbol{\theta}_0), ..., 2\hat{\gamma}_G(\mathbf{h}_H) - 2\gamma(\mathbf{h}_H, \boldsymbol{\theta}_0)\right) \xrightarrow{\mathcal{L}} N\left(\mathbf{0}, \boldsymbol{\Sigma}\right),$$

onde $\boldsymbol{\Sigma}$ é uma matriz diagonal tal que

$$\boldsymbol{\Sigma} = diag\{8\gamma(\mathbf{h}_1, \boldsymbol{\theta}_0)\mathrm{Var}[Q, F], ..., 8\gamma(\mathbf{h}_H, \boldsymbol{\theta}_0)\mathrm{Var}[Q, F]\}.$$

Demonstração: É consequência directa do **Teorema 6.4.15** e da independência dos $2\hat{\gamma}_G(\mathbf{h}_i)$, para $i = 1, ...H$.

□

Concluindo, sob certas condições de regularidade e num contexto de *IDA* ou de *MIDA*, o estimador usado na primeira fase da estimação pelo método dos múltiplos variogramas é consistente e tem uma distribuição assintótica normal.

De seguida, apresentam-se os resultados assintóticas relativos à segunda fase, ou seja, ao estimador de mínimos quadrados usuais. Os resultados de Lahiri *et al.* (2002) são directamente aplicáveis supondo as condições C.2, C.3 e C.4 seguintes.

C.2 Para qualquer $\varepsilon > 0$, existe um $\delta > 0$ tal que

$$\inf \left\{ \sum_{i=1}^{H} (2\gamma(\mathbf{h}_i, \boldsymbol{\theta}_1) - 2\gamma(\mathbf{h}_i, \boldsymbol{\theta}_2))^2 : \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| \geq \varepsilon \right\} > \delta;$$

C.3 para qualquer vector $\mathbf{h} \in \mathbb{R}^d$ fixo, $\sup\{\gamma(\mathbf{h}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\} < \infty$;

C.4 $\gamma(\mathbf{h}, \boldsymbol{\theta})$ tem derivadas parciais de ordem $s(\geq 0)$ em ordem a $\boldsymbol{\theta}$.

A condição C.2 serve para garantir a identificabilidade dos parâmetros do modelo de variograma. A identificabilidade, que já foi abordada na subsecção **6.4.2**, permite que, para qualquer $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, os variogramas $2\gamma(\mathbf{h}, \boldsymbol{\theta}_1)$ e $2\gamma(\mathbf{h}, \boldsymbol{\theta}_2)$ não sejam iguais em todos os $\mathbf{h}_i$, para $i = 1, ..., H$. Por outras palavras, os variogramas $2\gamma(\mathbf{h}, \boldsymbol{\theta}_1)$ e $2\gamma(\mathbf{h}, \boldsymbol{\theta}_2)$ têm que possuir valores diferentes em, pelo menos, um dos vectores $\mathbf{h}_i, i = 1, ..., H$, para que não existam soluções múltiplas por *OLS*. A condição C.3 verifica-se sem causar grandes restrições, pois basta que o espaço de parâmetros seja compacto – uma vez que o variograma é uma função contínua em $\boldsymbol{\theta} \in \Theta$, sendo Θ compacto, então a função é limitada. Finalmente a condição C.4 pode ser verificada directamente em cada modelo de variograma utilizado.

Para demonstrar a consistência do estimador de mínimos quadrados, apenas é necessária a continuidade de $2\gamma(\mathbf{h}, \boldsymbol{\theta})$ em $\boldsymbol{\theta}$. Isso consegue-se fazendo $s = 0$ na condição C.4. Para determinar a distribuição assintótica do estimador de mínimos quadrados, é necessária a diferenciabilidade de $2\gamma(\mathbf{h}, \boldsymbol{\theta})$, que se obtém tomando $s = 1$ em C.4. Comece-se então pela consistência do estimador de mínimos quadrados.

Teorema 6.4.18. Seja $2\hat{\gamma}(\mathbf{h}_i), i = 1, ..., H$, um estimador pontual do variograma e $\hat{\boldsymbol{\theta}}_n$ o estimador usual de mínimos quadrados. Assumam-se as condições C.2, C.3 e C.4 com $s = 0$. Se, para qualquer $i = 1, ..., H$, $2\hat{\gamma}(\mathbf{h}_i) \xrightarrow{\mathcal{P}} 2\gamma(\mathbf{h}_i, \boldsymbol{\theta}_0)$, então $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$.

Demonstração: Em Lahiri *et al.* (2002).

□

A partir do **Teorema 6.4.18** é imediato o seguinte resultado:

Corolário 6.4.19. Nas condições dos teoremas **6.4.15** e **6.4.18**, o estimador de mínimos quadrados, $\hat{\boldsymbol{\theta}}_n$, aplicado às estimativas pontuais do variograma, obtidas através do estimador Q_n de Genton, é consistente, isto é, $\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathcal{P}} \boldsymbol{\theta}_0$.

Demonstração: Resulta imediatamente da aplicação do **Teorema 6.4.18** atendendo a que, nas condições do **Teorema 6.4.15**, o estimador Q_n de Genton é consistente (*vide* **Corolário 6.4.16**).

□

A distribuição assintótica do estimador de mínimos quadrados, também foi estabelecida em Lahiri *et al.* (2002). Esse resultado utiliza vectores de derivadas parciais. Assim, é necessário considerar os vectores $g_j(\boldsymbol{\theta})$ das derivadas parciais de $\gamma(\mathbf{h}_1, \boldsymbol{\theta}), ...,$ $\gamma(\mathbf{h}_H, \boldsymbol{\theta})$ em ordem aos parâmetros $\theta_j \in \boldsymbol{\theta}$, ou seja, $g_j(\boldsymbol{\theta}_0) = ((\partial/\partial\theta_j)\gamma(\mathbf{h}_1, \boldsymbol{\theta}_0), ...,$ $(\partial/\partial\theta_j)\gamma(\mathbf{h}_H, \boldsymbol{\theta}_0)), 1 \leq j \leq q$. Além disso, agrupam-se os vectores das derivadas parciais numa matriz de dimensão $H \times q$ definida por $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) = [-2g_1(\boldsymbol{\theta}_0) \ ... \ -2g_q(\boldsymbol{\theta}_0)]$. Assim, é possível estabelecer o seguinte resultado:

Teorema 6.4.20. Seja $2\hat{\gamma}(\mathbf{h}_i), i = 1, ..., H$, um estimador pontual do variograma e $\hat{\boldsymbol{\theta}}_n$ o estimador usual de mínimos quadrados. Assuma-se que as condições C.2, C.3 e C.4 se verificam com $s = 1$ e que existe uma sucessão de constantes $\{a_n\}_{n\in\mathbb{N}}$, com $a_n \underset{n\to\infty}{\longrightarrow} +\infty,$, tal que

$$a_n \left(2\hat{\gamma}(\mathbf{h}_1) - 2\gamma(\mathbf{h}_1, \boldsymbol{\theta}_0), ..., 2\hat{\gamma}(\mathbf{h}_H) - 2\gamma(\mathbf{h}_H, \boldsymbol{\theta}_0)\right) \xrightarrow{\mathcal{L}} N_H\left(0, \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\right),$$

onde $\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$ é uma matriz não singular. Se a matriz $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ tiver característica q, então

$$a_n(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N_q\left(0, \boldsymbol{\Sigma}_\Gamma(\boldsymbol{\theta}_0)\right),$$

onde $\boldsymbol{\Sigma}_\Gamma(\boldsymbol{\theta}_0) = (\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0))^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)(\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)^{\mathrm{T}}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0))^{-1}$.

Demonstração: Em Lahiri *et al.* (2002).

□

O resultado anterior, garante a distribuição assintótica do estimador de mínimos quadrados, que é utilizado na segunda fase do método dos múltiplos variogramas. Repare-se que, para o estimador de múltiplos variogramas, como o número de estimativas pontuais do variograma é igual ao número de parâmetros a estimar, tem-se que $H = q$ e, portanto, $\mathbf{\Gamma}(\boldsymbol{\theta}_0)$ é uma matriz quadrada. Admitindo que $\mathbf{\Gamma}(\boldsymbol{\theta}_0)$ é não singular, é possível simplificar a matriz $\mathbf{\Sigma}_\Gamma(\boldsymbol{\theta}_0)$, obtendo-se que

$$\mathbf{\Sigma}_\Gamma(\boldsymbol{\theta}_0) = \mathbf{\Gamma}(\boldsymbol{\theta}_0)^{-1}\mathbf{\Sigma}(\boldsymbol{\theta}_0)(\mathbf{\Gamma}(\boldsymbol{\theta}_0)^{\mathrm{T}})^{-1}.$$

Corolário 6.4.21. Nas condições dos teoremas **6.4.15** e **6.4.20**, o estimador de mínimos quadrados $\hat{\boldsymbol{\theta}}_n$ aplicado às estimativas pontuais obtidas através do estimador Q_n de Genton, nos pontos $\mathbf{h}_1, ..., \mathbf{h}_H$ (com $H = q$), segue uma distribuição assintótica normal, isto é,

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{L}} N_q\left(0, \mathbf{\Sigma}_\Gamma(\boldsymbol{\theta}_0)\right),$$

onde $\mathbf{\Sigma}_\Gamma(\boldsymbol{\theta}_0) = \mathbf{\Gamma}(\boldsymbol{\theta}_0)^{-1}\mathbf{\Sigma}(\boldsymbol{\theta}_0)(\mathbf{\Gamma}(\boldsymbol{\theta}_0)^{\mathrm{T}})^{-1}$ e

$$\mathbf{\Sigma}(\boldsymbol{\theta}_0) = diag\{8\gamma(\mathbf{h}_1, \boldsymbol{\theta}_0)\mathrm{Var}[Q, F], ..., 8\gamma(\mathbf{h}_H, \boldsymbol{\theta}_0)\mathrm{Var}[Q, F]\}.$$

Demonstração: Resulta, imediatamente, da aplicação do **Teorema 6.4.20** atendendo a que, nas condições do **Teorema 6.4.15**, o estimador Q_n de Genton tem uma distribuição assintótica normal (*vide* **Corolário 6.4.17**).

□

Resumindo, as propriedades do estimador de múltiplos variogramas relativas à segunda etapa da estimação, dependem basicamente das propriedades dos estimadores utilizados na primeira etapa. Assim, como o estimador Q_n de Genton é consistente e assintoticamente normal, então o estimador de mínimos quadrados também o vai ser.

A terceira etapa do método consiste em repetir as duas etapas anteriores, produzindo um conjunto de curvas que constituem estimativas consistentes do variograma. Essas curvas são obtidas de forma independente uma vez que, na estimação de cada

curva, os vectores onde se obtêm as estimativas pontuais do variograma são seleccionados de forma independente. Resta então averiguar as propriedades da última etapa do método dos múltiplos variogramas.

A última etapa do estimador de múltiplos variogramas consiste em determinar a mediana das B estimativas de mínimos quadrados, as quais foram obtidas na terceira etapa da estimação. Isto é, repetindo-se a primeira e a segunda etapas B vezes para diferentes $\mathbf{h}_i, i = 1, ..., H$, obtém-se um conjunto de estimativas de mínimos quadrados $(\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_B)$. A estimativa final do método dos múltiplos variogramas é dada por

$$\tilde{\boldsymbol{\theta}}_B = \left(\text{Mediana}\{\hat{\theta}_{1,1}, ..., \hat{\theta}_{B,1}\}, ..., \text{Mediana}\{\hat{\theta}_{1,q}, ..., \hat{\theta}_{B,q}\}\right).$$

Como se mostrou no **Corolário 6.4.21**, quando a dimensão da amostra aumenta, cada $\hat{\boldsymbol{\theta}}_b = (\hat{\theta}_{b,1}, ..., \hat{\theta}_{b,q})$, para $b = 1, ..., B$, converge para uma distribuição assintótica normal com vector de média $\boldsymbol{\theta}_0$. Contudo, os $\hat{\boldsymbol{\theta}}_b$ não têm a mesma matriz de covariâncias. Considere-se então que n é suficientemente grande, para que seja possível utilizar a aproximação à distribuição normal. Neste contexto pretende-se averiguar como se comporta $\tilde{\boldsymbol{\theta}}_B$ à medida que B aumenta.

Note-se que a consistência e a distribuição assintótica normal da mediana são bem conhecidas quando as observações são *i.i.d.*. Por isso, os resultados que se seguem, correspondem a uma generalização desses casos.

Teorema 6.4.22. Sejam $\hat{\boldsymbol{\theta}}_b = \left(\hat{\theta}_{b,1}, ..., \hat{\theta}_{b,q}\right), b = 1, ..., B$, vectores aleatórios independentes e considere-se $\tilde{\boldsymbol{\theta}}_B = \left(\text{Mediana}\{\hat{\theta}_{1,1}, ..., \hat{\theta}_{B,1}\}, ..., \text{Mediana}\{\hat{\theta}_{1,q}, ..., \hat{\theta}_{B,q}\}\right)$. Se, para qualquer $b = 1, ..., B, \hat{\boldsymbol{\theta}}_b \sim N_q(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b)$, então

$$\tilde{\boldsymbol{\theta}}_B \xrightarrow[B\to\infty]{\mathcal{P}} \boldsymbol{\theta}_0.$$

Demonstração: Para cada $i = 1, ..., q$, fixo, seja

$$\bar{F}_{B,i} = \frac{1}{B}\sum_{b=1}^{B} F_{b,i},$$

onde $F_{b,i}$ é a *f.d.p.* de $\hat{\theta}_{b,i}$. Como $\hat{\boldsymbol{\theta}}_b \sim N_q(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_b)$, então $\hat{\theta}_{b,i} \sim N(\theta_{0,i}, \sigma^2_{b,i})$ e, a simetria da distribuição garante que

$$\bar{F}_{B,i}(\theta_{0,i}) = \frac{1}{B}\sum_{b=1}^{B} F_{b,i}(\theta_{0,i}) = \frac{1}{2}.$$

Por outro lado, dado que qualquer $F_{b,i}$ é uma função estritamente crescente, então $\bar{F}_{B,i}$ também o é. Por isso, para qualquer $\varepsilon > 0$, verifica-se que $1/2 < \bar{F}_{B,i}(\theta_{0,i} + \varepsilon) < 1$ e que $0 < \bar{F}_{B,i}(\theta_{0,i} - \varepsilon) < 1/2$. Consequentemente,

$$\sqrt{B}\left(\bar{F}_{B,i}(\theta_{0,i} + \varepsilon) - \frac{1}{2}\right) \underset{B}{\longrightarrow} \infty$$

e

$$\sqrt{B}\left(\frac{1}{2} - \bar{F}_{B,i}(\theta_{0,i} - \varepsilon)\right) \underset{B}{\longrightarrow} \infty.$$

Estas duas últimas condições, juntamente com o facto dos $\hat{\boldsymbol{\theta}}_b$ serem independentes, para $b = 1, ..., B$, permitem que se utilize um resultado de Mizera e Wellner (1998) (ver Teorema 1 do artigo) para concluir que

$$\text{Mediana}\{\hat{\theta}_{1,i}, ..., \hat{\theta}_{B,i}\} \underset{B}{\xrightarrow{\mathcal{P}}} \theta_{0,i}. \tag{6.4.10}$$

Como a consistência em (6.4.10) se verifica para qualquer $i = 1, ..., q$, então $\tilde{\boldsymbol{\theta}}_B$ tende em probabilidade para $\boldsymbol{\theta}_0$ quando $B \to \infty$.

□

O resultado anterior garante a consistência da última etapa do estimador de múltiplos variogramas. Portanto, o estimador de múltiplos variogramas é consistente.

Finalmente, falta concluir um resultado sobre a distribuição assintótica do estimador de múltiplos variogramas.

Teorema 6.4.23. Considere-se que se verificam as condições do **Teorema 6.4.22** e que $\lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} f_{b,i}(\theta_{0,i})$ existe e pertence a $\mathbb{R}^+$, onde $f_{b,i}$ representa a função densidade de probabilidade de $\hat{\theta}_{b,i}$, para todo o $i = 1, ..., q$. Então

$$\sqrt{B}\left(\text{Mediana}\{\hat{\theta}_{1,i}, ..., \hat{\theta}_{B,i}\} - \theta_{0,i}\right) \underset{B}{\xrightarrow{\mathcal{L}}} N\left(0, \left(2 \lim_{B\to\infty} \frac{1}{B} \sum_{b=1}^{B} f_{b,i}(\theta_{0,i})\right)^{-2}\right).$$

Demonstração: Para $i = 1, ..., q$ fixo, considerem-se as variáveis aleatórias $\hat{\theta}_{1,i}, ..., \hat{\theta}_{B,i}$, que, por hipótese, são independentes e seguem uma distribuição normal, com média $\theta_{0,i}$ e variância $\sigma^2_{b,i}$. A demonstração é desenvolvida tendo em conta um resultado de Koenker (2005), nomeadamente, utilizando o Teorema 4.1 do referido livro. Esse teorema assegura a distribuição assintótica normal do estimador dos coeficientes do modelo de regressão quantílica com erros heteroscedásticos, para um quantil τ genérico

$(0 < \tau < 1)$. Na presente demonstração, particulariza-se o resultado para $\tau = 0.5$, tomando $x_{b,i} = 1$, para $b = 1, ..., B$. Assim, para qualquer componente fixa i do parâmetro $\boldsymbol{\theta}$, obtém-se o modelo

$$\hat{\theta}_{b,i} = \theta_{0,i} + \epsilon_{b,i},$$

onde $\epsilon_{b,i}$ são os erros do modelo e $Q_{0.5}(\hat{\theta}_{b,i}|x_{b,i}) = \theta_{0,i}$, para qualquer $b = 1, ..., B$.

Se o modelo anterior estiver nas condições do Teorema 4.1 de Koenker (2005), então esse resultado permite provar o presente teorema. Assim, a demonstração prossegue comprovando que as condições necessárias à aplicação do Teorema 4.1 de Koenker (2005) se verificam.

A Condição A1 de Koenker (2005) verifica-se, uma vez que as varáveis aleatórias $\hat{\theta}_{b,i}$ seguem uma distribuição normal, a qual é absolutamente contínua e satisfaz $0 < f_{b,i}(\theta_{0,i}) < \infty$.

A Condição A2 de Koenker (2005) também se verifica uma vez que:

1. $\lim\limits_{B\to\infty} \dfrac{1}{B} \sum\limits_{b=1}^{B} x_{b,i} x_{b,i}^{\mathrm{T}} = [1]$, que é uma matriz definida positiva;

2. $\lim\limits_{B\to\infty} \dfrac{1}{B} \sum\limits_{b=1}^{B} f_{b,i}(\theta_{0,i}) x_{b,i} x_{b,i}^{\mathrm{T}} = \left[\lim\limits_{B\to\infty} \dfrac{1}{B} \sum\limits_{b=1}^{B} f_{b,i}(\theta_{0,i}) \right]$, que também é uma matriz definida positiva dado que, por hipótese, $\lim\limits_{B\to\infty} \dfrac{1}{B} \sum\limits_{b=1}^{B} f_{b,i}(\theta_{0,i}) > 0$;

3. por fim $\max\limits_{b=1,...,B} \dfrac{\|x_{b,i}\|}{\sqrt{B}} = \dfrac{1}{\sqrt{B}} \underset{B\to\infty}{\longrightarrow} 0$.

Dado que todas as condições do Teorema 4.1 de Koenker (2005) se verificam, fica concluída a demonstração.

□

O resultado anterior mostra que, quando o número de observações da amostra (n) e o número de variogramas considerados (B) são suficientemente grandes, então o estimador de múltiplos variogramas segue uma distribuição normal.

Resumindo, o estimador de múltiplos variogramas que se propôs é consistente e tem uma distribuição assintótica normal. É também um estimador robusto, porque é composto por um estimador robusto na primeira etapa – logo o estimador de múltiplos

variogramas possui uma função de influência limitada e um ponto de ruptura positivo. Por outro lado, o estimador de múltiplos variogramas proposto tem boa eficiência em modelos normais – repare-se que o estimador Q_n tem boa eficiência para estimar pontualmente o variograma em modelos normais e que, nas condições aqui consideradas, o estimador de mínimos quadrados é eficiente.

Concluindo, o estimador de múltiplos variogramas concilia boas propriedades de robustez com boa eficiência em modelos normais.

Capítulo 7

Exemplos de aplicação do estimador de múltiplos variogramas

O presente capítulo é dedicado à aplicação do estimador de múltiplos variogramas que foi proposto e estudado no capítulo anterior.

Assim, em primeiro lugar, apresenta-se um estudo de simulação, no qual se confronta o estimador de múltiplos variogramas com os métodos mais divulgados de estimação do variograma. O desempenho dos estimadores é avaliado considerando amostras de processos geoestatísticos geradas com diferentes modelos de variograma. Para cada modelo, estudam-se diversos graus de contaminação da amostra.

Na segunda parte do capítulo, analisa-se a aplicação do estimador de múltiplos variogramas a um conjunto de dados reais.

7.1 Estudo de simulação

Para efectuar um estudo comparativo do desempenho do estimador de múltiplos variogramas, consideraram-se outros estimadores do variograma, inseridos em programas informáticos especificamente vocacionados para as aplicações da Geoestatística. Assim, comparou-se o estimador de múltiplos variogramas com as seguintes alternativas:

- estimador tradicional – consiste na utilização do estimador de Matheron para estimar pontualmente o variograma, associado ao estimador de mínimos quadrados ponderados (WLS) para estimar os parâmetros do modelo de variograma. No método de WLS utilizam-se os pesos de Cressie (1993) que foram apresentados em (2.2.4);

- estimador Q_n em associação com *WLS* – o estimador de Matheron é substituído pelo estimador Q_n para estimar pontualmente o variograma; de seguida utiliza-se o *WLS*, com os pesos propostos por Cressie (1993), para estimar os parâmetros do modelo de variograma.

O estudo comparativo foi organizado do seguinte modo: para diferentes modelos de variograma, simularam-se amostras de processos geoestatísticos Gaussianos, quer sem contaminação, quer com diversos graus de contaminação. Em cada caso, foram contempladas amostras de diferentes dimensões. O critério de comparação foi o do erro quadrático médio.

Em particular, os modelos de variograma utilizados nas simulações foram os referidos em Genton (1998b). Assim, as amostras foram geradas de acordo com três modelos de variograma: o modelo esférico, o modelo exponencial e o modelo de potência, os quais são caracterizados, respectivamente, pelas expressões (1.3.7), (1.3.8) e (1.3.10). O modelo esférico e o modelo exponencial correspondem a processos estacionários de segunda ordem e são dos modelos de variograma mais utilizados nas aplicações da Geoestatística (*vide* Schabenberger e Gotway (2005)). Apesar de não ser tão utilizado na prática, o modelo de potência é o modelo de variograma mais divulgado de entre os que pertencem a processos apenas intrinsecamente estacionários.

Do ponto de vista da existência de parâmetros, é de relembrar que o modelo esférico pertence ao conjunto dos variogramas que têm amplitude, enquanto que o modelo exponencial não tem amplitude, mas tem amplitude prática; o modelo de potência não possui amplitude, nem amplitude prática. Por isso, esses três modelos são bons representantes dos diferentes variogramas que se utilizam usualmente em Geoestatística.

De acordo com o estudo de simulação publicado em Genton (1998b), geraram-se amostras com localizações dispostas aleatoriamente num segmento de recta com comprimento de 200 unidades. Note-se que, ao supor processos isotrópicos, o variograma do processo não depende da direcção considerada. Por isso, não há perda de generalidade ao efectuar o estudo do variograma através da simulação de pontos situados sobre uma recta.

Para gerar amostras de acordo com o modelo de variograma esférico ou com o modelo exponencial, usou-se o seguinte conjunto de valores de parâmetros:

- $\phi = 15$ (amplitude ou amplitude prática);
- $\tau^2 = 1$ (efeito de pepita);
- $\tau^2 + \sigma^2 = 3$ (patamar).

Quando as amostras foram simuladas de acordo com o modelo de potência apresentado em (1.3.10), considerou-se

- $\lambda = 0.5$;
- $\tau^2 = 0$ (efeito de pepita);
- $\theta = 2$.

Para observar o comportamento dos diferentes estimadores em relação ao aumento da dimensão da amostra, fez-se o mesmo estudo em amostras com 50 observações e em amostras com 200 observações.

Com o objectivo de avaliar a robustez do estimador de múltiplos variogramas, foram também mantidas as condições de contaminação estudadas por Genton. Deste modo, em cada amostra gerada, substituíram-se $\varepsilon\%$ de observações por observações independentes provenientes de uma distribuição $N(0, \sigma^2)$. A contaminação da amostra varia consoante as escolhas de ε e de σ. Neste estudo consideraram-se os seguintes casos:

A.1 Amostra sem contaminação;

A.2 Amostra com contaminação de 10% e $\sigma = 5$;

A.3 Amostra com contaminação de 20% e $\sigma = 5$;

A.4 Amostra com contaminação de 30% e $\sigma = 5$;

A.5 Amostra com contaminação de 10% e $\sigma = 10$;

A.6 Amostra com contaminação de 10% e $\sigma = 20$.

As contaminações A.2, A.3 e A.4, permitem averiguar como se comportam os estimadores quando o número de observações contaminadas aumenta. Nesse caso a distribuição

das observações contaminadas é sempre igual. As contaminações A.2, A.5 e A.6 permitem observar o comportamento dos estimadores, quando a variância da variável aleatória que gerou as observações contaminadas aumenta e o número de observações contaminadas permanece constante.

Simularam-se 1000 amostras de cada tipo de contaminação, para cada modelo de variograma considerado, quer para $n = 50$, quer para $n = 200$. Em cada caso, estimaram-se os variogramas através dos estimadores considerados, utilizando as regras práticas recomendadas em Journel e Huijbregts (1978), isto é, utilizando apenas pares de observações que estão separadas por uma distância inferior a metade da distância máxima entre localizações e considerando apenas as estimativas pontuais do variograma que foram obtidas com pelo menos 30 incrementos.

O estimador de múltiplos variogramas foi determinado com 250 repetições do processo iterativo, *i.e.*, a terceira etapa da estimação consistiu na obtenção de $B = 250$ estimativas do variograma, a partir das quais se obteve a estimativa final.

Em cada situação amostral, calculou-se o erro quadrático médio empírico de cada estimador, isto é,

$$EQME(\bar{\theta}_i^*) = \frac{1}{1000}\sum_{j=1}^{1000}\left(\bar{\theta}_i^{*(j)} - \theta_i\right)^2, i = 1, 2, 3,$$

onde $\bar{\theta}_i^{*(j)}$ é a estimativa do i-ésimo parâmetro do variograma, obtida na j-ésima amostra, e θ_i é o valor do i-ésimo parâmetro usado na simulação do modelo de variograma.

Os cálculos necessários ao estudo de simulação foram efectuados com o programa *R*. Para além da base do programa, utilizaram-se duas *packages* específicas – a *package geoR*, que serviu para simular as amostras e para trabalhar com as funções e estimadores tradicionais em Geoestatística; e a *package robustbase*, que permitiu obter as estimativas pelo estimador Q_n. Os resultados obtidos são apresentados nas tabelas que se seguem (Tabela 7.1 à Tabela 7.6).

Analisando os resultados obtidos, verifica-se que o estimador de múltiplos variogramas se comportou bastante melhor que os restantes estimadores considerados, uma vez que os seus erros quadráticos médios empíricos são quase sempre inferiores aos

Modelo esférico com $\phi = 15, \tau^2 = 1$ e $\sigma^2 = 2$ ($n = 50$).				
Contaminação	Método	EQME($\hat{\phi}$)	EQME($\hat{\tau}^2$)	EQME($\hat{\tau}^2 + \hat{\sigma}^2$)
A.1	Math. *WLS*	402.615	0.318	1.375
	Q_n *WLS*	376.581	0.380	1.717
	Mult. variog.	141.389	0.253	1.113
A.2	Math. *WLS*	476.080	7.590	12.667
	Q_n *WLS*	415.593	1.732	6.355
	Mult. variog.	166.621	1.489	5.549
A.3	Math. *WLS*	428.559	24.019	31.948
	Q_n *WLS*	387.574	6.060	16.966
	Mult. variog.	214.242	6.124	17.098
A.4	Math. *WLS*	352.776	52.121	66.025
	Q_n *WLS*	413.158	19.210	40.767
	Mult. variog.	231.812	19.287	41.931
A.5	Math. *WLS*	467.990	98.230	216.365
	Q_n *WLS*	478.496	2.885	19.395
	Mult. variog.	222.453	2.527	14.027
A.6	Math. *WLS*	407.576	1522.922	3773.975
	Q_n *WLS*	526.673	3.956	35.813
	Mult. variog.	238.036	3.721	24.565

Tabela 7.1: Erros quadráticos médios empíricos dos estimadores dos parâmetros do modelo esférico, para amostras de dimensão 50. Os estimadores considerados foram o estimador de Matheron com mínimos quadrados ponderados (Math. *WLS*), o estimador Q_n com mínimos quadrados ponderados (Q_n *WLS*) e o estimador de múltiplos variogramas (Mult. variog.).

Modelo esférico com $\phi = 15, \tau^2 = 1$ e $\sigma^2 = 2$ ($n = 200$).				
Contaminação	Método	EQME($\hat{\phi}$)	EQME($\hat{\tau}^2$)	EQME($\hat{\tau}^2 + \hat{\sigma}^2$)
A.1	Math. *WLS*	289.850	0.103	0.539
	Q_n *WLS*	369.682	0.117	0.663
	Mult. variog.	40.021	0.043	0.462
A.2	Math. *WLS*	395.172	7.104	6.506
	Q_n *WLS*	284.331	0.714	2.619
	Mult. variog.	34.998	0.574	2.206
A.3	Math. *WLS*	459.959	26.146	22.218
	Q_n *WLS*	207.917	4.168	9.085
	Mult. variog.	56.484	3.903	8.844
A.4	Math. *WLS*	457.074	56.074	47.923
	Q_n *WLS*	264.367	14.796	23.327
	Mult. variog.	95.673	14.560	24.859
A.5	Math. *WLS*	529.829	98.081	109.556
	Q_n *WLS*	238.009	1.421	7.410
	Mult. variog.	43.019	1.150	6.411
A.6	Math. *WLS*	266.372	783.580	1883.447
	Q_n *WLS*	312.207	2.069	15.237
	Mult. variog.	68.785	1.694	13.087

Tabela 7.2: Erros quadráticos médios empíricos dos estimadores dos parâmetros do modelo esférico, para amostras de dimensão 200. Os estimadores considerados foram o estimador de Matheron com mínimos quadrados ponderados (Math. *WLS*), o estimador Q_n com mínimos quadrados ponderados (Q_n *WLS*) e o estimador de múltiplos variogramas (Mult. variog.).

Modelo exponencial com $\phi = 15, \tau^2 = 1$ e $\sigma^2 = 2$ ($n = 50$).				
Contaminação	Método	EQME($\hat{\phi}$)	EQME($\hat{\tau}^2$)	EQME($\hat{\tau}^2 + \hat{\sigma}^2$)
A.1	Math. *WLS*	262.530	0.703	1.413
	Q_n *WLS*	308.059	0.707	1.981
	Mult. variog.	229.484	0.606	1.708
A.2	Math. *WLS*	265.100	6.542	12.771
	Q_n *WLS*	268.897	2.150	7.206
	Mult. variog.	156.347	1.771	7.538
A.3	Math. *WLS*	379.720	20.554	37.345
	Q_n *WLS*	366.607	5.983	19.684
	Mult. variog.	119.687	5.435	22.088
A.4	Math. *WLS*	238.865	34.716	67.712
	Q_n *WLS*	315.745	15.054	44.619
	Mult. variog.	188.709	13.666	47.211
A.5	Math. *WLS*	266.978	60.517	216.529
	Q_n *WLS*	292.258	3.135	19.928
	Mult. variog.	217.932	2.562	17.234
A.6	Math. *WLS*	474.325	1634.102	4190.424
	Q_n *WLS*	275.142	4.583	42.758
	Mult. variog.	175.427	3.505	28.391

Tabela 7.3: Erros quadráticos médios empíricos dos estimadores dos parâmetros do modelo exponencial, para amostras de dimensão 50. Os estimadores considerados foram o estimador de Matheron com mínimos quadrados ponderados (Math. *WLS*), o estimador Q_n com mínimos quadrados ponderados (Q_n *WLS*) e o estimador de múltiplos variogramas (Mult. variog.).

Modelo exponencial com $\phi = 15, \tau^2 = 1$ e $\sigma^2 = 2$ ($n = 200$).				
Contaminação	Método	EQME($\hat{\phi}$)	EQME($\hat{\tau}^2$)	EQME($\hat{\tau}^2 + \hat{\sigma}^2$)
A.1	Math. *WLS*	119.994	0.440	0.421
	Q_n *WLS*	120.974	0.465	0.489
	Mult. variog.	12.459	0.268	0.332
A.2	Math. *WLS*	372.786	5.733	6.897
	Q_n *WLS*	186.597	0.952	2.773
	Mult. variog.	40.670	0.668	2.456
A.3	Math. *WLS*	372.444	18.030	22.778
	Q_n *WLS*	174.527	3.576	9.285
	Mult. variog.	57.973	3.317	10.925
A.4	Math. *WLS*	261.862	35.840	46.995
	Q_n *WLS*	238.163	11.204	23.999
	Mult. variog.	56.102	11.404	33.688
A.5	Math. *WLS*	308.278	60.192	108.701
	Q_n *WLS*	169.519	1.655	7.892
	Mult. variog.	18.632	1.225	7.345
A.6	Math. *WLS*	165.507	867.612	1941.534
	Q_n *WLS*	155.750	2.179	14.351
	Mult. variog.	30.530	1.602	13.270

Tabela 7.4: Erros quadráticos médios empíricos dos estimadores dos parâmetros do modelo exponencial, para amostras de dimensão 200. Os estimadores considerados foram o estimador de Matheron com mínimos quadrados ponderados (Math. *WLS*), o estimador Q_n com mínimos quadrados ponderados (Q_n *WLS*) e o estimador de múltiplos variogramas (Mult. variog.).

Modelo de potência com $\lambda = 0.5; \tau^2 = 0$ e $\theta = 2$ ($n = 50$).				
Contaminação	Método	EQME($\hat{\lambda}$)	EQME($\hat{\tau}^2$)	EQME($\hat{\theta}$)
A.1	Math. *WLS*	0.179	2.023	3.321
	Q_n *WLS*	0.155	1.930	4.300
	Mult. variog.	0.142	0.692	1.791
A.2	Math. *WLS*	0.256	32.349	14.467
	Q_n *WLS*	0.140	5.754	7.175
	Mult. variog.	0.114	2.283	4.110
A.3	Math. *WLS*	0.294	70.779	37.819
	Q_n *WLS*	0.158	17.903	20.317
	Mult. variog.	0.102	4.880	14.710
A.4	Math. *WLS*	0.326	95.819	72.934
	Q_n *WLS*	0.210	45.964	45.535
	Mult. variog.	0.113	7.877	40.920
A.5	Math. *WLS*	0.230	89.821	96.212
	Q_n *WLS*	0.142	6.840	13.233
	Mult. variog.	0.122	2.815	7.102
A.6	Math. *WLS*	0.315	516.942	1255.600
	Q_n *WLS*	0.190	11.687	22.464
	Mult. variog.	0.132	4.043	10.993

Tabela 7.5: Erros quadráticos médios empíricos dos estimadores dos parâmetros do modelo de potência, para amostras de dimensão 50. Os estimadores considerados foram o estimador de Matheron com mínimos quadrados ponderados (Math. *WLS*), o estimador Q_n com mínimos quadrados ponderados (Q_n *WLS*) e o estimador de múltiplos variogramas (Mult. variog.).

Modelo de potência com $\lambda = 0.5; \tau^2 = 0$ e $\theta = 2$ ($n = 200$).				
Contaminação	Método	EQME($\hat{\lambda}$)	EQME($\hat{\tau}^2$)	EQME($\hat{\theta}$)
A.1	Math. *WLS*	0.149	2.020	1.803
	Q_n *WLS*	0.123	1.564	1.727
	Mult. variog.	0.110	0.761	1.298
A.2	Math. *WLS*	0.210	26.772	10.925
	Q_n *WLS*	0.091	3.664	4.189
	Mult. variog.	0.070	1.571	2.795
A.3	Math. *WLS*	0.262	74.173	29.839
	Q_n *WLS*	0.104	10.686	14.206
	Mult. variog.	0.064	2.824	10.487
A.4	Math. *WLS*	0.278	121.081	57.475
	Q_n *WLS*	0.127	26.118	39.212
	Mult. variog.	0.078	4.542	33.827
A.5	Math. *WLS*	0.267	112.022	63.227
	Q_n *WLS*	0.086	4.263	7.737
	Mult. variog.	0.061	1.512	5.083
A.6	Math. *WLS*	0.301	584.321	773.772
	Q_n *WLS*	0.093	5.461	10.472
	Mult. variog.	0.070	2.312	6.930

Tabela 7.6: Erros quadráticos médios empíricos dos estimadores dos parâmetros do modelo de potência, para amostras de dimensão 200. Os estimadores considerados foram o estimador de Matheron com mínimos quadrados ponderados (Math. *WLS*), o estimador Q_n com mínimos quadrados ponderados (Q_n *WLS*) e o estimador de múltiplos variogramas (Mult. variog.).

dos outros estimadores. É de salientar que os resultados encontrados com o estimador de múltiplos variogramas são melhores do que os do estimador tradicional, mesmo em amostras que não têm contaminação. Em amostras contaminadas, o estimador de múltiplos variogramas continua a ter um desempenho bastante bom, quase sempre melhor do que o do outro estimador robusto considerado.

Quando se analisa apenas os resultados dos modelos de variograma de processos estacionários de segunda ordem, é de salientar que o estimador de múltiplos variogramas provoca uma melhoria significativa na estimação da amplitude. Tal como Genton (1998b) mostrou, a amplitude é o parâmetro mais importante dos dois processos estacionários de segunda ordem que aqui foram considerados, uma vez que é o único parâmetro que tem impacto nos pesos da *Krigagem* tradicional. O efeito de pepita e o patamar apenas influenciam a variância da *krigagem*. Sendo assim, de seguida faz-se uma análise detalhada da estimação da amplitude.

A Figura 7.1 mostra como variam os erros quadráticos médios empíricos dos estimadores da amplitude, em função da percentagem de observações contaminadas na amostra. Nesse caso, as observações contaminadas foram geradas com variáveis aleatórias *i.i.d.*, com distribuição normal com média zero e desvio padrão igual a 5.

Repare-se que, em qualquer uma das situações, o erro quadrático médio empírico do estimador de múltiplos variogramas é sempre inferior aos dos outros estimadores. Portanto, no que diz respeito à estimação da amplitude, o estimador de múltiplos variogramas tem uma vantagem nítida em relação aos outros estimadores considerados. Analisando agora apenas os outros estimadores e tal como era de esperar, o estimador de Matheron comporta-se melhor para amostras sem contaminação, enquanto que o estimador robusto se revela melhor quando aumenta a percentagem de observações contaminadas na amostra.

Para completar a análise do desempenho do estimador de múltiplos variogramas no que respeita à robustez/contaminação da amostra, tem interesse analisar o comportamento do estimador à medida que se aumenta o desvio padrão da distribuição que gera as observações contaminadas. Na Figura 7.2 apresentam-se os valores dos erros quadráticos médios empíricos dos estimadores da amplitude, para diferentes valores do desvio padrão. Neste caso, a percentagem de observações contaminadas foi de 10%,

para qualquer um dos valores do desvio padrão considerados.

Verificou-se que os erros quadráticos médios empíricos do estimador de múltiplos variogramas são muito menores do que os dos restantes estimadores, para todos os valores de desvio padrão analisados. Portanto, o estimador de múltiplos variogramas comporta-se bastante bem quando o desvio padrão das observações contaminadas aumenta.

Na Figura 7.2 pode ainda observar-se que, no modelo esférico, o estimador de Matheron tem um erro quadrático médio empírico menor do que o do estimador Q_n com *WLS*, quando o desvio padrão das observações contaminadas é igual a 20. Neste caso, esperar-se-ia um comportamento melhor do estimador robusto. Uma justificação possível para este facto, resulta do desvio padrão das observações contaminadas ser muito elevado, o que faz com que as estimativas do variograma devolvidas pelo estimador de Matheron, sejam muito afectadas. De facto, a forte variabilidade introduzida pela contaminação faz com que o estimador de Matheron não "detecte"a estrutura de dependência do processo – as estimativas encontradas parecem resultar de variáveis não correlacionadas. Como a amplitude indica o valor a partir do qual as observações são não correlacionadas, o estimador devolve estimativas da amplitude muito próximas de zero. Assim, a variância do estimador de Matheron diminui de tal forma, que faz com que o erro quadrático médio empírico diminua consideravelmente.

No entanto, nas condições do parágrafo anterior e apesar do estimador de Matheron apresentar menor erro quadrático médio empírico do que o estimador Q_n com *WLS*, é de realçar que o primeiro é globalmente pior estimador do que este último, uma vez que ignora a estrutura de dependência. Esta afirmação vem ilustrada na Figura 7.3. Os processos que dão origem às estimativas representadas nessa figura, foram gerados com amplitude $\phi = 15$, valor esse que se pretendia estimar. Mas, observando os diagramas de extremos e quartis da distribuição empírica das estimativas da amplitude para cada estimador, torna-se evidente que só os estimadores robustos resistiram à contaminação introduzida.

Na Figura 7.3 também se pode observar que, em todos os estimadores considerados, as estimativas da amplitude parecem ser provenientes de uma distribuição assimétrica. Este facto também se verificou nas amostras sem contaminação e nas estimativas dos

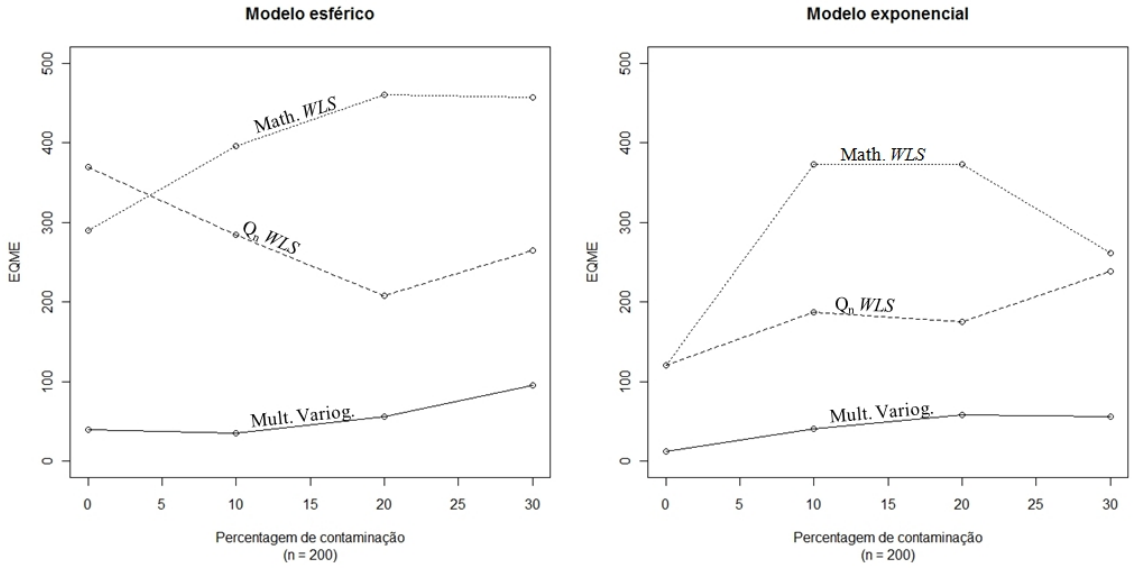

Figura 7.1: Erros quadráticos médios empíricos dos estimadores da amplitude em função da percentagem de observações contaminadas na amostra.

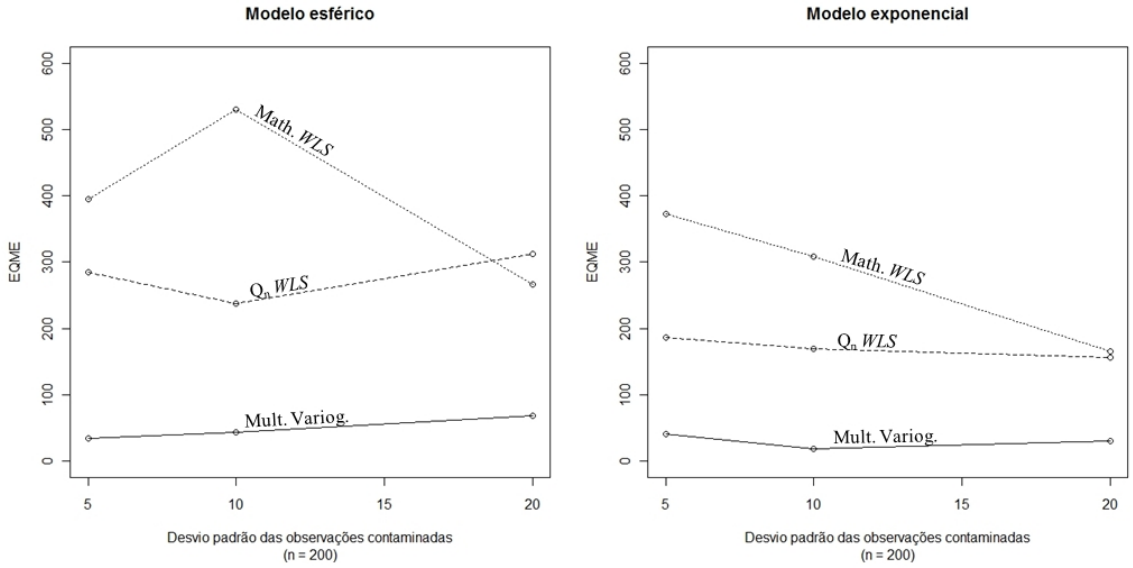

Figura 7.2: Erros quadráticos médios empíricos dos estimadores da amplitude em função do desvio padrão das observações contaminadas da amostra.

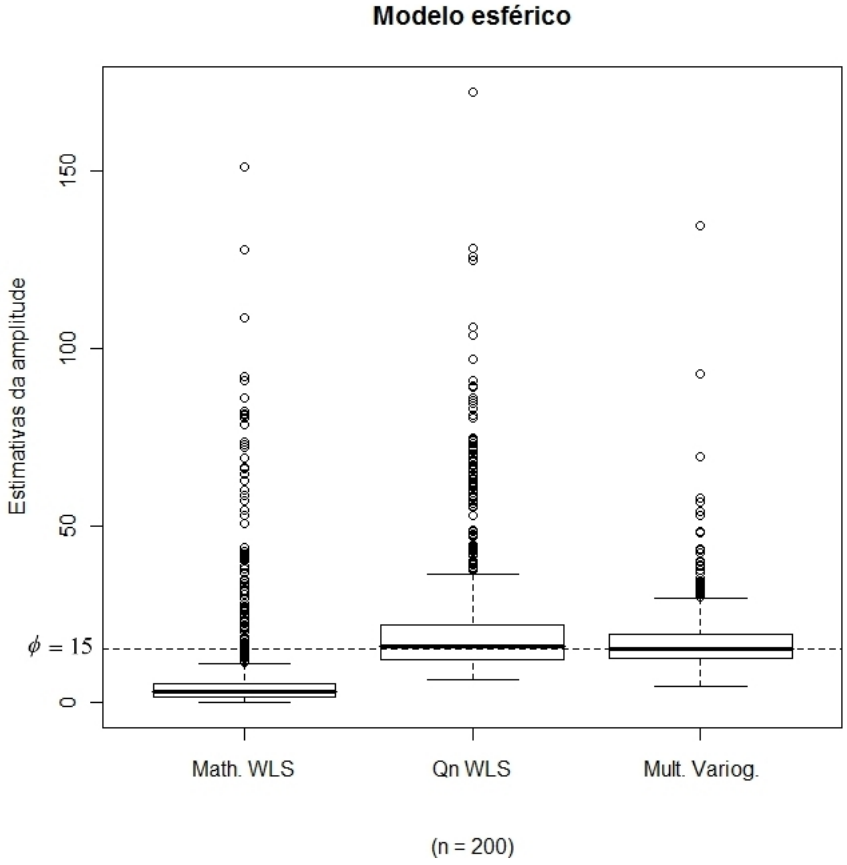

Figura 7.3: Diagramas de extremos e quartis das estimativas da amplitude, para amostras com 10% de observações contaminadas, geradas a partir de uma distribuição $N(0, 20^2)$. A linha a tracejado representa o valor da amplitude usado na simulação.

outros parâmetros. Por isso, a distribuição dos estimadores para esta dimensão amostral (n = 200), ainda não deve ser aproximada pela distribuição normal. De facto, a convergência em lei para a distribuição normal, que foi demonstrada no capítulo anterior para o estimador de múltiplos variogramas, é muito lenta, pelo que é necessária uma dimensão amostral bastante grande para que seja possível utilizá-la. Esta constatação também é válida para o estimador de Matheron com *WLS*. Lahiri *et al.* (2002) mostraram que o estimador de Matheron com *WLS* também converge em lei para a distribuição normal. Porém, a Figura 7.3 e o estudo de simulação efectuado, revelam que a convergência desse estimador para a distribuição normal também é bastante lenta.

Os comentários anteriores referem-se a simulações de processos estacionários de segunda ordem. Considere-se agora a análise de processos intrinsecamente estacionários, através da simulação de processos com variograma pertencente ao modelo de potência. Neste caso, o parâmetro mais importante é o expoente λ, uma vez que é este que determina a forma do variograma. Repare-se que, se $\lambda < 1$, a curva do variograma tem a concavidade voltada para baixo; se $\lambda > 1$, a curva do variograma tem a concavidade voltada para cima; por fim, se $\lambda = 1$, obtém-se o modelo de variograma linear.

A Figura 7.4 mostra como variam os erros quadráticos médios empíricos dos estimadores de λ, quer em função da percentagem de contaminação presente na amostra, quer em função do desvio padrão das observações contaminadas. Quando varia a percentagem de observações contaminadas, elas são sempre provenientes de uma distribuição normal com média 0 e desvio padrão igual a 5. Quando varia o desvio padrão das observações contaminadas, a amostra contém sempre 10% de contaminação.

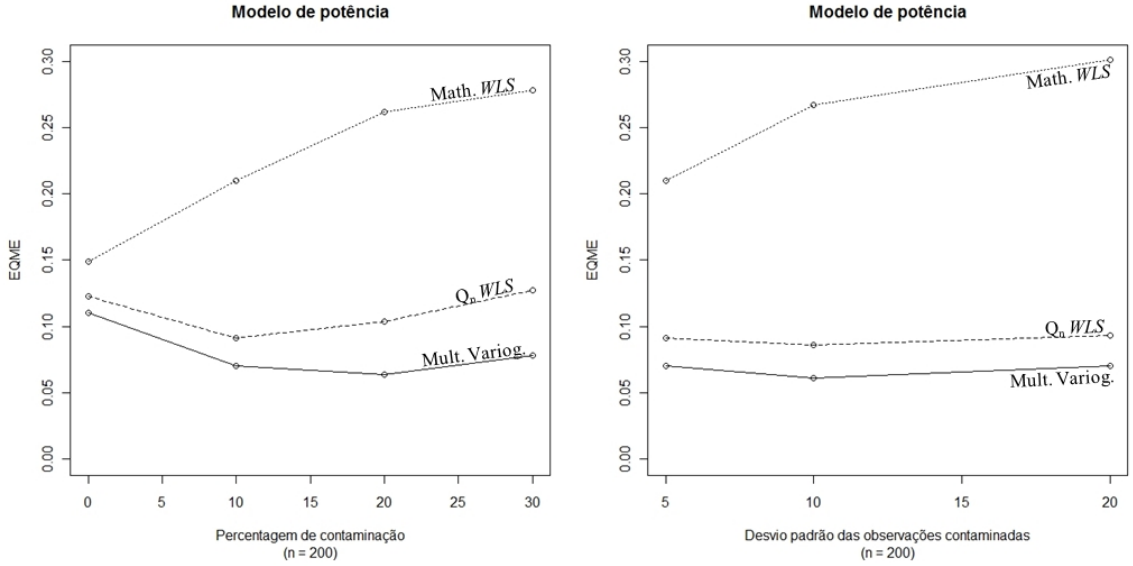

Figura 7.4: Erros quadráticos médios empíricos dos estimadores de λ do modelo de potência. À esquerda estão expressos em função da percentagem de observações contaminadas e à direita em função do desvio padrão dessas observações.

Como se pode observar, o estimador de múltiplos variogramas tem o erro quadrático médio empírico menor do que os restantes, em todas as situações consideradas. Sendo

assim, o estimador de múltiplos variogramas é aquele que devolve as melhores estimativas de λ. Neste caso, o estimador Q_n com *WLS* também revela um bom comportamento. No entanto, o erro quadrático médio empírico do estimador Q_n com *WLS* de λ, nunca chega a ser menor do que o do estimador de múltiplos variogramas.

A análise do estudo efectuado, incidiu sobre os parâmetros mais importantes de cada modelo de variograma considerado e permitiu concluir que, nos casos considerados, o estimador de múltiplos variogramas tem boas propriedades. Não se apresenta uma análise idêntica em relação aos restantes parâmetros, pelo facto do estudo de simulação ter conduzido às mesmas conclusões, o que se confirma pelos valores apresentados desde a Tabela 7.1 até à Tabela 7.6.

Concluindo, de um modo geral, o estimador de múltiplos variogramas produz melhores estimativas do que os restantes estimadores considerados. Em particular, e apesar de ser robusto, o estimador de múltiplos variogramas revelou-se melhor do que as alternativas, mesmo em amostras onde não existe contaminação.

Assim, os resultados obtidos neste estudo de simulação salientam o bom desempenho do estimador de múltiplos variogramas e confirmam as propriedades estudadas no capítulo anterior.

7.2 Análise de um conjunto de dados reais

Nesta secção analisa-se um conjunto de dados reais através dos métodos de estimação do variograma considerados na secção anterior. Os dados utilizados são de acesso livre e fazem parte da *package geoR* do *software R.* Assim, para aceder a este conjunto de dados, basta aceder à *package geoR* que os contém sob o nome *soil.*

Como o próprio nome indica, os dados *soil* representam um conjunto de características químicas, medidas na superfície do solo. O conjunto contém observações recolhidas em 250 localizações, dispostas ao longo de uma grelha regular de 10×25 pontos e, em cada localização considerada, mediram-se as características químicas que se pretendiam avaliar. Mais informação sobre este conjunto de dados pode ser encontrada em Bassoi (1994).

De entre as diversas variáveis registadas neste conjunto de dados, focou-se o estudo da distribuição da quantidade de potássio presente no solo. Assim, em cada localização $\mathbf{s} \in D$, a observação $Z(\mathbf{s})$ representa a quantidade de potássio aí encontrada. A escolha da variável foi motivada pelo facto de existirem potenciais observações atípicas na amostra.

Nas Figuras 7.5 e 7.6 observam-se duas representações gráficas da amostra da quantidade de potássio presente no solo.

A Figura 7.5 é uma ilustração da amostra, onde se representam os valores da quantidade de potássio relativos a cada localização amostrada, como se o terreno fosse visto de cima. Quanto maior e mais escuro for o círculo representado, maior é a quantidade de potássio presente na localização. Analisando a figura, é possível visualizar duas localizações onde a quantidade de potássio é muito elevada – a observação que se encontra na linha 3 e na coluna 13 é a que se torna mais evidente, uma vez que contrasta bastante com os valores observados nas localizações vizinhas; a segunda observação mais relevante, que corresponde à linha 7 e coluna 20, já não apresenta uma diferença tão acentuada em relação às observações vizinhas. Outro aspecto que vale a pena notar, diz respeito ao topo esquerdo da figura. Repare-se que existe aí um conjunto de observações onde a quantidade de potássio é pequena, quando comparada com as restantes observações da amostra.

Os comentários referidos no parágrafo anterior, confirmam-se na análise da Figura 7.6, onde a amostra é representada num gráfico tridimensional. Os dois picos bastante pronunciados correspondem às duas observações já identificadas, enquanto o vale à esquerda corresponde às observações onde a quantidade de potássio é pequena. Todas estas observações que foram salientadas são potenciais observações atípicas e, por isso, sugerem que seja utilizado um estimador robusto para estimar o variograma.

Como foi salientado no **Capítulo 5**, ao analisar um processo geoestatístico deve-se começar por verificar se não existem motivos para rejeitar a hipótese de estacionaridade da média. Para tal observe-se a Figura 7.7, onde se podem visualizar as projecções ortogonais das observações da amostra, ao longo das direcções definidas pela grelha das localizações, *i.e.*, ao longo da direcção do vector $\vec{i} = (1, 0)$ e do vector $\vec{j} = (0, 1)$. Observando a figura, parecem não existir motivos significativos para rejeitar a existência

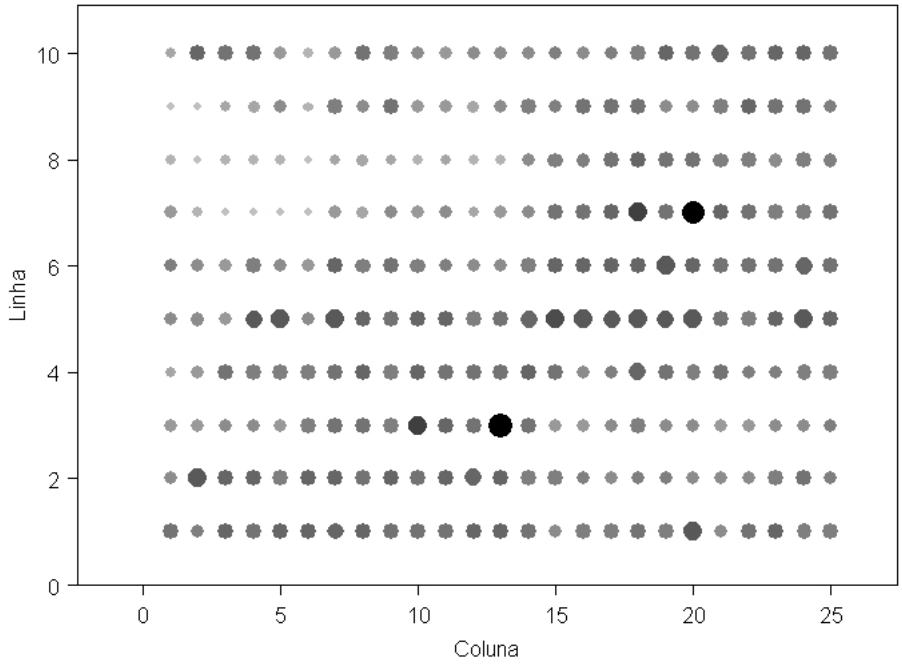

Figura 7.5: Representação da amostra da quantidade de potássio. Quanto maior e mais escuro for o círculo, maior é a quantidade de potássio presente nessa localização.

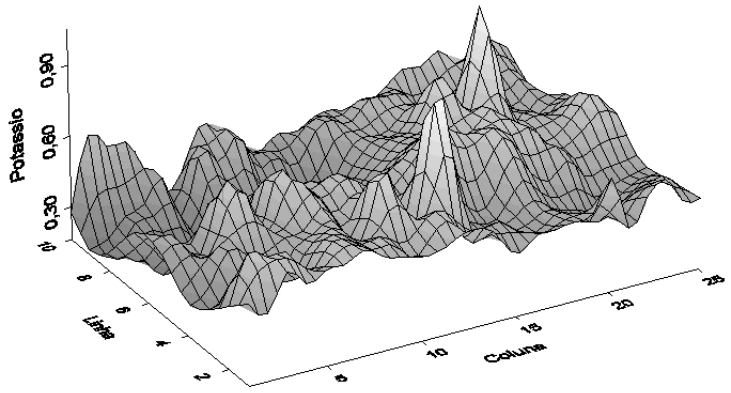

Figura 7.6: Representação tridimensional da amostra da quantidade de potássio presente no solo.

de estacionaridade da média, uma vez que se nota uma tendência, mas não muito pronunciada. No entanto, decidiu-se efectuar um teste à estacionaridade da média, de acordo com a secção **5.3**.

Neste caso, testou-se a existência de estacionaridade da média contra a hipótese alternativa mais simples – a da existência de uma tendência linear – traduzida por uma recta de regressão.

O teste foi efectuado ao longo das duas direcções dos vectores $\vec{i}$ e $\vec{j}$. Deste modo, estimou-se o declive das rectas de regressão em ambas as direcções, usando o estimador LAD e um estimador-MM. Na direcção de $\vec{i}$, *i.e.*, ao longo das linhas da grelha, obtiveram-se $\hat{\beta}_{LAD} = 0.004$ e $\hat{\beta}_{MM} = 0.005$. Na direcção de $\vec{j}$ obtiveram-se $\hat{\beta}_{LAD} = -0.008$ e $\hat{\beta}_{MM} = -0.009$. As estimativas foram calculadas através do *software R*. Utilizou-se a *package quantreg* para determinar $\hat{\beta}_{LAD}$ e a *package robustbase* para determinar $\hat{\beta}_{MM}$. Ambas as *packages* foram usadas com as opções existentes por defeito. Ainda na Figura 7.7, pode-se observar as rectas de regressão estimadas com o estimador LAD (linha a tracejado) e com o estimador-MM (linha a cheio).

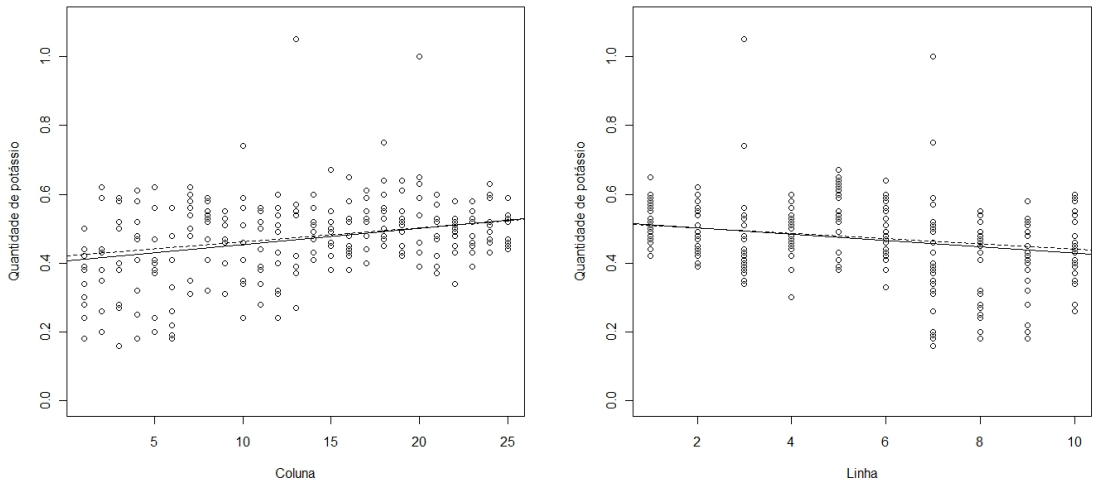

Figura 7.7: Observações da quantidade de potássio em função das colunas (à esquerda) e das linhas (à direita) e rectas de regressão estimadas pelo LAD (linha a tracejado) e por um estimador-MM (linha a cheio).

A Figura 7.7 mostra a semelhança entre as duas rectas estimadas, em cada um dos casos. Também é possível observar claramente as duas observações discordantes que já

tinham sido detectadas anteriormente.

Para prosseguir com o teste à estacionaridade da média, é necessário estimar o desvio padrão do LAD e do estimador-MM nas direcções de $\vec{i}$ e $\vec{j}$, respectivamente, $\hat{s}_{\hat{\beta}_{\vec{i}}}$ e $\hat{s}_{\hat{\beta}_{\vec{j}}}$.

Para tal, utilizou-se a metodologia *bootstrap SFBB* apresentada na subsecção **5.3.2**. Como, numa análise preliminar dos dados, a amplitude de $Z(\mathbf{s})$ foi estimada em valores próximos das 5 unidades, considerou-se que $Z(\mathbf{s})$ é um processo m-dependente, com $m = 5$. Deste modo, para preservar a estrutura de dependência dentro de cada um dos blocos *bootstrap* a reamostrar, os blocos foram formados com $L = m/2 = 2.5$ (*vide* expressão (5.3.3)). Repare-se que, neste caso, como só existem 10 linhas, não é razoável considerar $L = m = 5$, tal como foi feito na subsecção **5.3.4**, e formar blocos de 10×10 observações, pois quando se procede desse modo, obtêm-se poucos blocos para reamostrar.

Através da metodologia *bootstrap*, obtiveram-se estimativas iguais para o desvio padrão de ambos os estimadores, LAD e estimador-MM, para qualquer uma das direcções $\vec{i}$ e $\vec{j}$, em particular, $\hat{s}_{\hat{\beta}_{\vec{i}}} \approx 0.003$ e $\hat{s}_{\hat{\beta}_{\vec{j}}} \approx 0.010$.

Efectuando o teste a um nível de significância $\alpha = 5\%$, o ponto crítico da região explicitada em (5.3.7) é aproximadamente igual a $q_{1,2} = 2.24$. Consequentemente, para o estimador LAD, verificou-se que o valor observado da estatística do teste não pertence à região crítica, pois,

$$\left|\frac{\hat{\beta}_{LAD}}{\hat{s}_{\hat{\beta}_{\vec{i}}}}\right| = \left|\frac{0.004}{0.003}\right| \approx 1.33 < 2.24 \wedge \left|\frac{\hat{\beta}_{LAD}}{\hat{s}_{\hat{\beta}_{\vec{j}}}}\right| = \left|\frac{-0.008}{0.010}\right| \approx 0.8 < 2.24.$$

Obteve-se a mesma conclusão para o estimador-MM, uma vez que

$$\left|\frac{\hat{\beta}_{MM}}{\hat{s}_{\hat{\beta}_{\vec{i}}}}\right| = \left|\frac{0.005}{0.003}\right| \approx 1.67 < 2.24 \wedge \left|\frac{\hat{\beta}_{MM}}{\hat{s}_{\hat{\beta}_{\vec{j}}}}\right| = \left|\frac{-0.009}{0.010}\right| \approx 0.9 < 2.24.$$

Portanto, quer com o estimador LAD, quer com o estimador-MM conclui-se que, com base na amostra recolhida e com o nível de significância de 5%, não houve motivos significativos para rejeitar a hipótese nula. Assim, assumiu-se que o processo $Z(\mathbf{s})$ tem estacionaridade da média.

Passando à estimação do variograma, em primeiro lugar é necessário escolher o

modelo de variograma que é mais adequado a este processo. Uma análise preliminar permitiu assumir que o variograma deste processo é isotrópico.

A Figura 7.8 mostra um conjunto de estimativas pontuais do semivariograma, que foram obtidas através do estimador Q_n de Genton (1998a). As estimativas pontuais

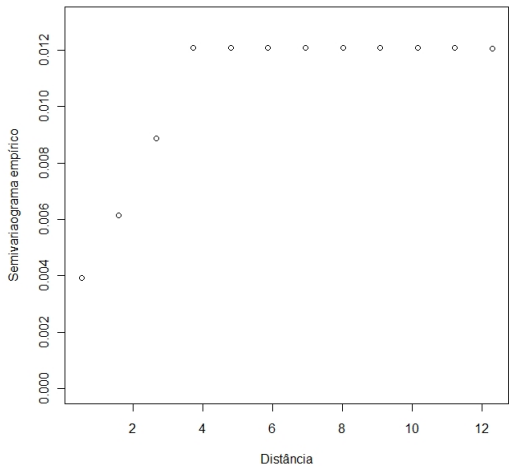

Figura 7.8: Estimativas pontuais do semivariograma obtidas através do estimador Q_n.

do semivariograma sugerem que se opte por um modelo de variograma com amplitude e patamar. De entre os modelos com esses parâmetros, a representação gráfica aponta para a escolha de um modelo de tenda. No entanto, o modelo de tenda só é válido em processos de domínio contido em $\mathbb{R}$. Como não é o caso, o modelo de tenda não pode ser aqui utilizado.

Desse modo, seguidamente considera-se a modelação por um modelo circular e por um modelo esférico. Considerando que o variograma do processo em estudo é isotrópico e que pertence à família de variogramas de modelo circular, estimaram-se os parâmetros do variograma através dos métodos utilizados na secção anterior. Admitindo um modelo esférico, procedeu-se do mesmo modo. As Figuras 7.9 e 7.10 representam, respectivamente, as estimativas do semivariograma com modelo circular e as estimativas do semivariograma com modelo esférico, obtidos com os diferentes métodos de estimação.

Nas figuras pode-se ver que as estimativas obtidas pelo estimador de Matheron

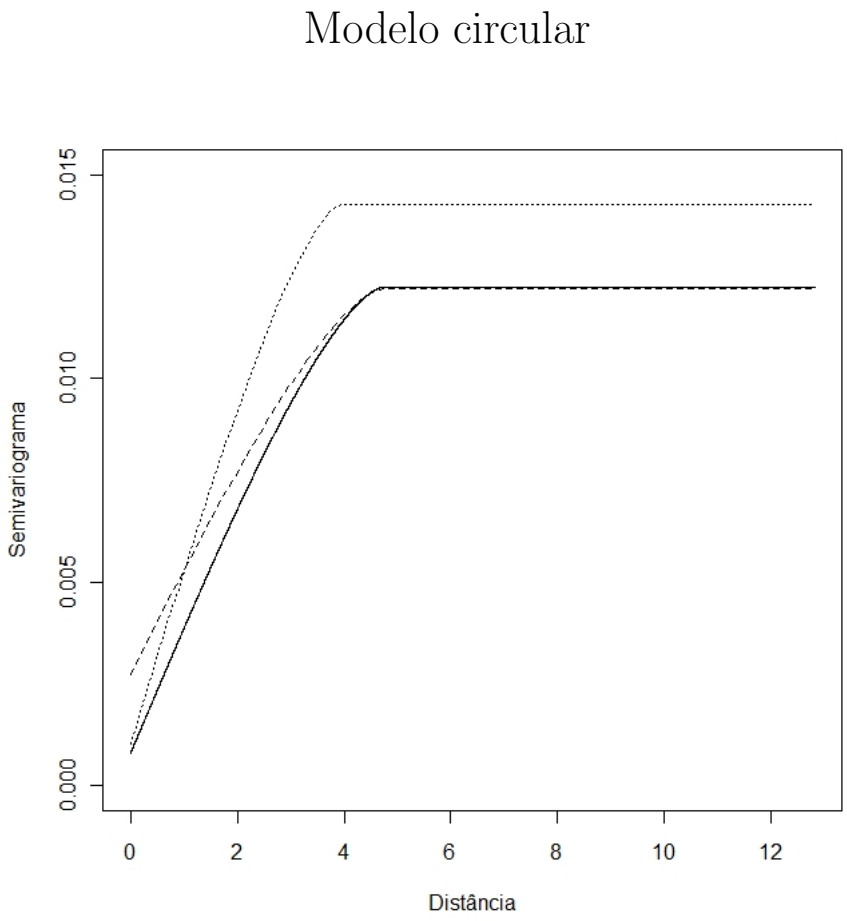

Figura 7.9: Estimativas do semivariograma com modelo circular, obtidas com os seguintes métodos de estimação: estimador de Matheron com *WLS* (linha a picotado), estimador Q_n com *WLS* (linha a tracejado) e estimador de múltiplos variogramas (linha a cheio).

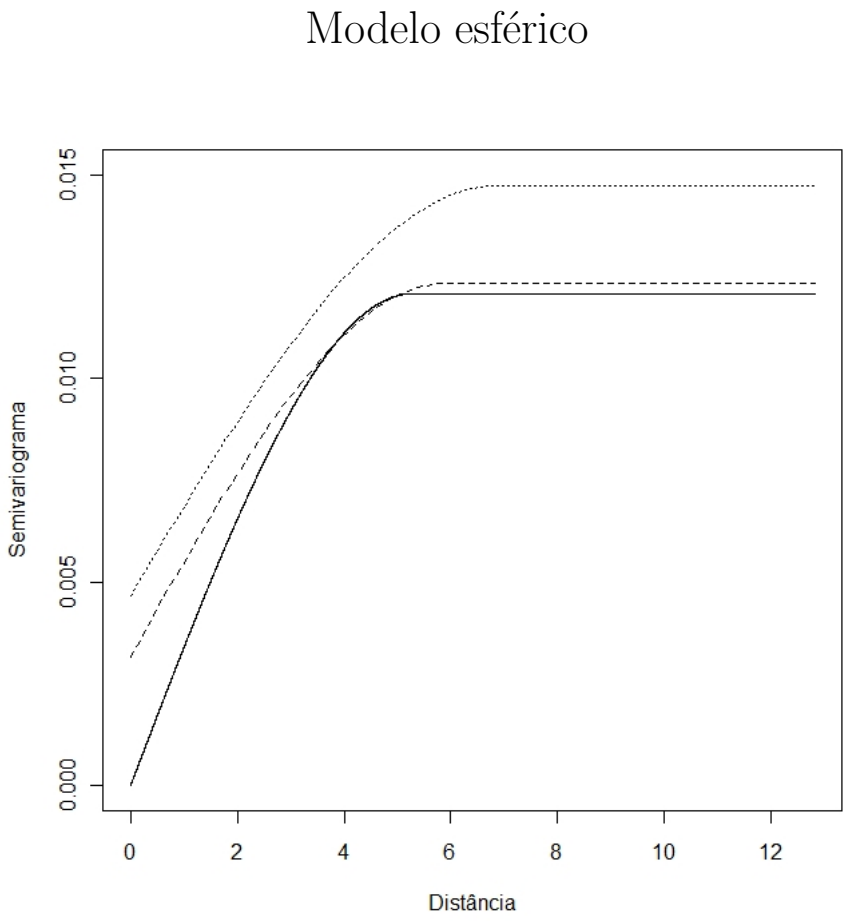

Figura 7.10: Estimativas do semivariograma com modelo esférico, obtidas com os seguintes métodos de estimação: estimador de Matheron com *WLS* (linha a picotado), estimador Q_n com *WLS* (linha a tracejado) e estimador de múltiplos variogramas (linha a cheio).

com *WLS* se afastam das encontradas com os restantes estimadores. Este facto poderá dever-se à existência das observações atípicas. Quanto aos dois estimadores robustos, eles parecem comportar-se de forma semelhante.

As estimativas obtidas encontram-se nas tabelas 7.7 e 7.8, respectivamente, para o modelo circular e para o modelo esférico.

Modelo circular

Parâmetro	Math. *WLS*	Q_n *WLS*	Mult. Variog.
ϕ	3,9405	4,7183	4,7183
τ^2	0,0010	0,0027	0.0008
$\tau^2 + \sigma^2$	0,0142	0,0122	0,0121

Tabela 7.7: Estimativas dos parâmetros do modelo de variograma circular, obtidas pelos seguintes métodos: estimador de Matheron com *WLS*, estimador Q_n com *WLS* e estimador de múltiplos variogramas.

Modelo esférico

Parâmetro	Math. *WLS*	Q_n *WLS*	Mult. Variog.
ϕ	6,8779	5,8954	5,2756
τ^2	0,0047	0,0032	0
$\tau^2 + \sigma^2$	0,0148	0,0124	0,0121

Tabela 7.8: Estimativas dos parâmetros do modelo de variograma esférico, obtidas pelos seguintes métodos: estimador de Matheron com *WLS*, estimador Q_n com *WLS* e estimador de múltiplos variogramas.

Uma vez que se trata de dados reais, os valores dos parâmetros dos modelos são desconhecidos, pelo que é impossível efectuar uma análise semelhante à da secção anterior. Assim, resta verificar se as estimativas encontradas pelo estimador de múltiplos variogramas são coerentes com a alternativa obtida pelo outro estimador robusto. Efectivamente, não há grande diferença entre as estimativas da amplitude e do patamar produzidas pelos dois estimadores robustos. Por outro lado, em relação à estimação do efeito de pepita, verificou-se que o estimador de múltiplos variogramas apresenta estimativas menores do que as obtidas usando o estimador Q_n com *WLS*, em qualquer um dos modelos considerados.

Para concluir este capítulo, é de salientar que o estimador de múltiplos variogramas foi de encontro às expectativas, confirmando, na prática, as boas propriedades de robustez e de eficiência em modelos normais, as quais eram o objectivo principal do presente trabalho.

Conclusões

Neste trabalho descreve-se o estudo que foi efectuado sobre estimação robusta do variograma. Procurou-se um método de estimação com boas propriedades de robustez mas que, simultaneamente, possuísse boa eficiência em processos geoestatísticos Gaussianos univariados e que, para além disso, fosse simples de utilizar em aplicações a dados reais. Como resultado da investigação conduzida, recomenda-se um novo método, que se designou por estimador de múltiplos variogramas.

Em termos gerais, a tese consiste em três partes: a primeira parte consiste num conjunto de noções e resultados fundamentais para o desenvolvimento do novo estimador; a segunda parte contém essencialmente propostas e resultados originais; e a terceira parte exemplifica o cálculo das estimativas de múltiplos variogramas.

Como a modelação por processos geoestatísticos assume a hipótese da estacionaridade da média do processo, é importante poder validar esta hipótese. Neste trabalho propôs-se um teste estatístico para confirmar a estacionaridade da média. O teste foi construído recorrendo a estimadores-MM, os quais foram propostos por Yohai (1987). Para efectuar o referido teste, foi necessário aproximar a distribuição da estatística do teste através de metodologias de reamostragem. Assim, adaptou-se a metodologia *bootstrap* a processos espaciais com a devida precaução, de modo a preservar a dependência existente entre as observações originais. De acordo com o esquema de reamostragem sugerido, verificou-se que, sob condições de regularidade, a estatística de teste que foi utilizada possui uma distribuição assintótica normal, o que facilita a aplicação do teste por parte dos utilizadores da Geoestatística.

Admitindo a existência de estacionaridade da média, prosseguiu-se com a estimação do variograma. Na proposta apresentada neste trabalho, sugere-se que o processo de estimação do variograma consista em mais fases do que as duas etapas de estimação que

são tradicionais. Assim, na primeira etapa recomenda-se a utilização de um estimador robusto, altamente resistente, para obter as estimativas pontuais do variograma – por exemplo, pode-se utilizar o estimador Q_n como foi proposto em Genton (1998a). Na segunda fase utilizam-se as estimativas pontuais obtidas, para estimar os parâmetros desconhecidos do modelo de variograma. Para tal, usa-se um método que assegura boas propriedades de eficiência, recorrendo ao facto de, sob certas condições, o método dos mínimos quadrados simples poder ser tão eficiente como o método dos mínimos quadrados generalizados. Tais condições admitem que o número de estimativas pontuais do variograma seja igual ao número de parâmetros desconhecidos do modelo. É, por isso, da maior importância, que se utilize um estimador robusto na primeira etapa da estimação. Seguidamente, melhora-se a eficiência global do método, tirando partido dos meios informáticos actualmente disponíveis – concretamente, repetem-se as duas fases anteriores, sucessivamente, variando os pontos onde são calculadas (de forma robusta) as estimativas pontuais do variograma. Deste modo, gera-se um conjunto de múltiplos variogramas, produzidos a partir de uma única amostra observada. Por fim, na quarta etapa, obtém-se a estimativa final do variograma através de uma medida de tendência central das estimativas que pertencem ao conjunto de múltiplos variogramas. Nesta última etapa também se propôs a utilização de um estimador bastante robusto, como é o caso da mediana, para obter as estimativas centrais dos parâmetros do modelo.

A estimação dos parâmetros do modelo pressupõe a sua identificabilidade. Contudo, demonstrou-se que, ao usar estimativas pontuais na primeira fase da estimação, a identificabilidade nem sempre está assegurada. Foi possível estabelecer condições que contornam o problema, garantindo a existência de uma solução única no processo de optimização que conduz ao cálculo das estimativas. Os resultados demonstrados também são válidos quando se usam processos tradicionais de estimação.

Como os estimadores do variograma são definidos implicitamente, é necessário que as estimativas sejam determinadas através de procedimentos computacionais. Assim, o presente trabalho incluiu uma componente computacional relevante. Houve o cuidado de utilizar *software* disponível e de livre acesso, como é o caso do programa *R* e das suas *packages*, para que os procedimentos propostos estejam devidamente testados e possam ser facilmente utilizados por todos os interessados.

As conclusões alcançadas ao longo da investigação foram claramente satisfatórias. Por um lado, quando se aplicou o novo método a um conjunto de dados reais já publicados na literatura, os resultados obtidos foram próximos dos encontrados por outros procedimentos robustos. Por outro lado, em estudos de simulação, concluiu-se que o método recomendado neste trabalho, conduziu a melhores resultados do que os restantes procedimentos considerados, quer em situações com contaminação, quer em situações sem contaminação.

Note-se que a obtenção de estimativas pelo método de múltiplos variogramas, exige um esforço computacional maior do que pelos métodos tradicionais. No entanto, é de realçar que os custos computacionais adicionais não são relevantes face às vantagens evidenciadas pelo estimador que se propôs. Para isso, contribuem a evolução das capacidades informáticas e a facilidade de acesso a técnicas da estatística computacional.

O trabalho deixa em aberto uma série de questões que poderão ser objecto de futuras pesquisas. A título de exemplo, será interessante estudar, futuramente, a utilização dos estimadores-MM na primeira fase de estimação do variograma, em substituição do estimador Q_n de Genton (1998a) – essa substituição poderá vir a aumentar a eficiência do método de múltiplos variogramas. Por outro lado, o estudo dos próprios estimadores-MM continua a ser um tópico de grande actualidade. Finalmente, na sequência da estimação do variograma, segue-se frequentemente o processo de *krigagem*. Certamente que será da maior utilidade investigar métodos robustos de *krigagem*, num complemento do estudo até agora efectuado.

Para finalizar, espera-se que este trabalho possa vir a ser um contributo para o avanço e a divulgação dos métodos robustos, e em particular, da sua aplicação à Geoestatística.

Bibliografia

Bassoi, L. H. (1994). *Nitrato no solo e acumulo de N pelo milho (Zea mays L.) fertirrigado*, Tese de Doutoramento, Universidade de São Paulo, Brasil.

Beaton, A. E. e Tukey, J. W. (1974). The fitting of power series, meaning polynomials, illustrated on hand-spectroscopic data, *Technometrics* **16**: 147–185.

Box, G. (1953). Non-normality and tests on variances, *Biometrika* **40**: 318–335.

Bradley, R. (2005). Properties of strong mixing conditions. a survey and some open questions., *Probability Surveys* **2**: 107–144.

Carlstein, E. (1992). Resampling technics for stationary time series: some recent developments, *in* D. Brillinger, P. Caines, J. Geweke, E. Parzen, M. Rosenblatt e M. Taqqu (eds), *New Directions in Time Series Analysis. Part 1*, Springer, pp. 75–85.

Casella, G. e Berger, R. L. (2002). *Statistical Inference – Second Edition*, Duxbury Advanced Series, U.S.A.

Christofides, T. e Mavrikiou, P. (2003). Central limit theorem for dependent multidimensionally indexed random variables, *Statistics and Probability Letters* **63**: 67–78.

Cressie, N. (1993). *Statistics for Spatial Data – Revised Edition*, John Wiley and Sons Inc., U.S.A.

Cressie, N. e Hawkins, D. M. (1980). Robust estimation of the variogram, I, *Journal of the International Association for Mathematical Geology* **12**: 115–125.

Croux, C. e Rousseeuw, P. (1992). Time-efficient algorithms for two highly robust estimators of scale, *in* Y. Dodge e J. Whittaker (eds), *Computational Statistics*, Vol. 1, Physica-Verlag, Heidelberg, pp. 411–428.

Davison, A. e Hinkley, D. (1997). *Bootstrap Methods and their Application*, Cambridge University Press.

Efron, B. (1979). Bootstrap methods: another look at the jackknife, *The Annals of Statistics* **7**: 1–26.

Fitzenberger, B. (1997). The moving blocks bootstrap and robust inference for linear least squares and quantile regressions, *Journal of Econometrics* **82**: 235–287.

Genton, M. G. (1998a). Highly robust variogram estimation, *Mathematical Geology* **30**(2): 213–221.

Genton, M. G. (1998b). Variogram fitting by generalized least squares using an explicit formula for the covariance structure, *Mathematical Geology* **30**(4): 323–345.

Genton, M. G. (2001). Robustness problems in the analysis of spatial data, *in* M. Moore (ed.), *Spatial Statistics: Methodological Aspects and Applications*, Springer-Verlag, New York, pp. 21–38.

Gorsich, D. e Genton, M. (2000). Variogram model selection via nonparametric derivative estimation, *Journal of International Association for Mathematical Geology* **32**(3): 249–270.

Hall, P. (1985). Resampling a coverage pattern, *Stochastic Processes and their Applications* **20**: 231–246.

Hampel, F. R. (1968). *Contributions to the Theory of Robust Estimation*, Tese de Doutoramento, Universidade da California, Berkeley.

Hampel, F., Ronchetti, E., Rousseeuw, P. e Stahel, W. (1986). *Robust Statistics: the approach based on influence functions*, John Wiley and Sons Inc., New York.

Hart, J. F. (1954). Central tendency in areal distributions, *Economic geography* **30**: 48–59.

Hawkins, D. M. (1981). A cusum for a scale parameter, *Journal of Quality Technology* **13**: 228–231.

Huber, P. (1981). *Robust Statistics*, John Wiley and Sons Inc., New York.

Huber, P. J. (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics* **35**: 73–101.

Journel, A. G. e Huijbregts, C. J. (1978). *Mining Geostatistics*, Academic Press, London.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations, *The Annals of Statistics* **17**(3): 1217–1241.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press, Cambridge.

Koenker, R. e Basset, G. (1978). Regression quantiles, *Econometrica* **46**: 33–50.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*, Springer, New York.

Lahiri, S. N., Lee, Y. e Cressie, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters, *Journal of Statistical Planning and Inference* **103**: 65–85.

Maglione, D. e Diblasi, A. (2001). Choosing a valid model for the variogram of an isotropic spatial process, *2001 Annual Conference of International Association for Mathematical Geology*.

Mardia, K. e Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression, *Biometrika* **71**: 135–146.

Maronna, R. A., Martin, R. D. e Yohai, V. J. (2006). *Robust Statistics – Theory and Methods*, John Wiley and Sons Inc., Great Britain.

Matheron, G. (1962). Traite de geostatistique appliquee, tome I, Vol. 14 of *Memoires du Bureau de Recherches Geologiques et Minieres*, Editions Technip, Paris.

Matheron, G. (1963). Principles of geostatistics, *Economic Geology* **58**: 1246–1266.

Matheron, G. (1971). The theory of regionalized variables and its applications, Vol. 5 of *Cahiers du Centre de Morphologie Mathematique*, Fontainebleau, France.

Mizera, I. e Wellner, J. A. (1998). Necessary and sufficient conditions for weak consistency of the median of independent but not identically distributed random variables, *The Annals of Statistics* **26**(2): 672–691.

Politis, D. e Romano, J. (1992). A circular block resampling procedure for stationary data, *in* H. Lepage e L. Billard (eds), *Exploring the Limits of Bootstrap*, John Wiley and Sons Inc., pp. 263–270.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ribeiro Jr., P. J. e Diggle, P. J. (2001). geoR: A package for geostatistical analysis, *R-News* **1**(2): 15–18.

Rousseeuw, P. J. e Croux, C. (1993). Alternatives to the median absolute deviation, *Journal of the American Statistical Association* **88**(424): 1273–1283.

Rousseeuw, P. J. e Yohai, V. J. (1984). Robust regression by means of S-estimators, *Robust and Nonlinear Time Series Analysis*, Vol. 26 of *Lecture Notes in Statistics*, Springer, New York, pp. 256–272.

Salibian-Barrera, M. (2006). The asymptotics of MM-estimators for linear regression with fixed designs, *Metrika* **63**: 283–294.

Schabenberger, O. e Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*, Chapman & Hall/CRC, U.S.A.

Serfling, R. J. (1984). Generalized L-, M- and R-statistics, *The Annals of Statistics* **12**(1): 76–86.

Shakesby, R., Coelho, C., Schnabel, S., Keizer, J., Clarke, M., Contador, J. L., Walsh, R., Ferreira, A. e Doerr, S. (2002). Ranking as a potential methodology for assessing relative erosion risk and its application to dehesas and montados in spain and portugal, *Land Degradation and Development* **13**: 129–140.

Soares, A. (2000). *Geoestatística para as ciências da terra e do ambiente*, IST Press, Lisboa.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression, *The Annals of Statistics* **15**(2): 642–656.

Zimmerman, D. L. e Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding kriging predictors, *Technometrics* **33**: 77–91.